ADVANCES IN INTELLIGENT AND *SOFT COMPUTING* 166

David C. Wyld
Jan Zizka
Dhinaharan Nagamalai (Eds.)

Advances in Computer Science, Engineering and Applications

Proceedings of the second International Conference on Computer Science, Engineering and Applications (ICCSEA 2012), May 25-27, 2012, New Delhi, India, Volume 1

Springer

Advances in Intelligent and Soft Computing 166

Editor-in-Chief

Prof. Janusz Kacprzyk
Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warsaw
Poland
E-mail: kacprzyk@ibspan.waw.pl

For further volumes:
http://www.springer.com/series/4240

David C. Wyld, Jan Zizka,
and Dhinaharan Nagamalai (Eds.)

Advances in Computer Science, Engineering and Applications

Proceedings of the Second International Conference on Computer Science, Engineering and Applications (ICCSEA 2012), May 25–27, 2012, New Delhi, India, Volume 1

Springer

Editors
David C. Wyld
Southeastern Louisiana University
Hammond
USA

Dhinaharan Nagamalai
Wireilla Net Solutions PTY Ltd
Melbourne
Australia

Jan Zizka
Mendel University
Brno
Czech Republic

ISSN 1867-5662 e-ISSN 1867-5670
ISBN 978-3-642-30156-8 e-ISBN 978-3-642-30157-5
DOI 10.1007/978-3-642-30157-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2012937233

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

The Second International Conference on Computer Science, Engineering and Applications (ICCSEA-2012) was held in Delhi, India, during May 25–27, 2012. ICCSEA-2012 attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West. The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The ICCSEA-2012 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the conference. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer-review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically. All these efforts undertaken by the Organizing and Technical Committees led to an exciting, rich and a high quality technical conference program, which featured high-impact presentations for all attendees to enjoy, appreciate and expand their expertise in the latest developments in computer Science and Engineering research.

In closing, ICCSEA-2012 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research.

It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come. We hope that you will benefit from the fine papers from the ICCSEA-2012 conference that are in this volume and will join us at the next ICCSEA conference.

David C. Wyld
Jan Zizka
Dhinaharan Nagamalai

Organization

General Chairs

David C. Wyld	Southeastern Louisiana University, USA
Michal Wozniak	Wroclaw University of Technology, Poland
Henrique Joao Lopes Domingos	University of Lisbon, Portugal

Steering Committee

Jose Enrique Armendariz-Inigo	Universidad Publica de Navarra, Spain
John Karamitsos	University of the Aegean, Samos, Greece
Dhinaharan Nagamalai	Wireilla Net Solutions PTY LTD, Australia
Chih-Lin Hu	National Central University, Taiwan
Salah M. Saleh Al-Majeed	University of Essex, United Kingdom
Jan Zizka	SoNet/DI, FBE, Mendel University in Brno, Czech Republic

Program Committee Members

A. Kannan	K.L.N. College of Engineering, India
A.P. Sathish Kumar	PSG Institute of Advanced Studies, India
Abdul Aziz	University of Central Punjab, Pakistan
Abdul Kadhir Ozcan	Karatay University, Turkey
Ahmed Nada	Al-Quds University, Palestinian
Alejandro Regalado Mendez	Universidad del Mar - México, USA
Ali M.	University of Bradford, United Kingdom
Ali Maqousi	Petra University, Jordan
Anand Sharma	MITS -Rajasthan, India
Andy Seddon	Asia Pacific Institute of Information Technology, Malaysia
Anjan K.	RVCE–Bangalore, India
Ankit Thakkar	Nirma University, India

Anthony Atayero	Covenant University, Nigeria
Ashok Kumar Das	IIT Hyderabad, India
B. Srinivasan	Monash University, Australia
Balasubramanian K.	KLefke European University, Cyprus
Balasubramanian Karuppiah	MGR University, India
Beatrice Cynthia Dhinakaran	Hannam University, South Korea
Bela Genge	European Commission Joint Research Centre, Belgium
Bobby Barua	Ahsanullah University of Science and Technology, Bangladesh
Bong-Han Kim	Chongju University, South Korea
Boo-Hyung Lee	KongJu National University, South Korea
Brajesh Kumar Kaushik	Indian Institute of Technology, India
Carlos E. Otero	University of South Florida Polytechnic, USA
Ch.V. Rama Rao	Gudlavalleru Engineering College, India
Charalampos Z. Patrikakis	National Technical University of Athens, Greece
Chih-Lin Hu	National Central University, Taiwan
Chin-Chih Chang	Chung Hua University,Taiwan
Cho Han Jin	Far East University, South Korea
Danda B. Rawat	Old Dominion University, USA
Khamish Malhotra	University of Glamorgan, UK
M. Rajarajan	City University, UK
Mohammad Momani	University of Technology Sydney, Australia
Raja Kumar M.	National Advanced IPv6 Center (NAv6), Universiti Sains Malaysia
Salman Abdul Moiz	Centre for Development of Advanced Computing, India
Carlos E. Otero	The University of Virginia's College at Wise, USA
Wojciech Mazurczyk	Warsaw University of Technology, Poland
David C. Wyld	Southeastern Louisiana University, India
David W. Deeds	Shingu College, South Korea
Debasis Giri	Haldia Institute of Technology, India
Dhinaharan Nagamalai	Wireilla Net Solutions PVT ltd, Australia
Dimitris Kotzinos	Technical Educational Institution of Serres, Greece
E. Martin	University of California, Berkeley, USA
Emmanuel Bouix	iKlax Media, France
Ermatita Zuhairi	Sriwijaya University, Indonesia
Farhat Anwar	International Islamic University, Malaysia
Firkhan Ali Bin Hamid Ali	Universiti Tun Hussein Onn Malaysia, Malaysia
Ford Lumban Gaol	University of Indonesia
Ghalem Belalem	University of Oran, Algeria
Giovanni Schembra	University of Catania, Italy
Girija Chetty	University of Canberra, Australia
Gomathi Kandasamy	Avinashilingam Deemed University for Women, India

H.V. Ramakrishnan	Dr. MGR University, India
Hao-En Chueh	Yuanpei University, Taiwan
Henrique Joao Lopes Domingos	University of Lisbon, Portugal
Ho Dac Tu	Waseda University, Japan
Hoang Huu Hanh	Hue University, Vietnam
Hussein Al-Bahadili	Petra University, Jordan
Hwangjun Song	Pohang University of Science and Technology, South Korea
Intisar Al-Mejibli	University of Essex, United Kingdom
J.K. Mandal	University of Kalyani, India
Jacques Demerjian	Communication & Systems, Homeland Security, France
Jae Kwang Lee	Hannam University, South Korea
Jan Zizka	SoNet/DI, FBE, Mendel University in Brno, Czech Republic
Jeong-Hyun Park	Electronics Telecommunication ResearchInstitute, South Korea
Jeyanthy N.	VIT University, India
Jifeng Wang	University of Illinois at Urbana Champaign, USA
Jivesh Govil	Cisco Systems Inc., USA
Johann Groschdl	University of Bristol, UK
John Karamitsos	University of the Aegean, Greece
Johnson Kuruvila	Dalhousie University, Canada
Jose Enrique Armendariz-Inigo	Universidad Publica de Navarra, Spain
Jungwook Song	Konkuk University, South Korea
K.P. Thooyamani	Bharath University, India
Kai Xu	University of Bradford, United Kingdom
Kamalrulnizam Abu Bakar	Universiti Teknologi Malaysia, Malaysia
Khamish Malhotra	University of Glamorgan, UK
Krishnamurthy E.V.	ANU College Of Engg & Computer Science, Austraila
Krzysztof Walkowiak	Wroclaw University of Technology, Poland
Krzysztof Walkowiak	Wroclaw University of Technology, Poland
Lu Yan	University of Hertfordshire, UK
Lus Veiga	Technical University of Lisbon, Portugal
M. Aqeel Iqbal	FUIEMS, Pakistan
Mahesh Goyani	G.H. Patel College of Engineering and Technology, India
Maragathavalli P.	Pondicherry Engineering College, India
Marco Folli	University of Pavia, Italy
Marco Roccetti	Universty of Bologna , Italy
Martin A.	Pondicherry University, India
Massimo Esposito	ICAR-CNR, Italy
Michal Wozniak	Wroclaw University of Technology, Poland
Mohammad Ali Jabreil Jamali	Islamic Azad University, Iran
Mohammad Zaidul Karim	Daffodil International University, Bangladesh

Mohsen Sharifi	Iran University of Science and Technology, Iran
Moses Ekpenyong	University of Uyo, Nigeria
Muhammad Sajjadur Rahim	University of Rajshahi, Bangladesh
Murugan D.	Manonmaniam Sundaranar University, India
N. Kaliammal	NPR College of Engg & Tech, India
N. Krishnan	Manonmaniam Sundaranar University, India
Nabendu Chaki	University of Calcutta, India
Naohiro Ishii	Aichi Institute of Technology, Japan
Nasrollah M. Charkari	Tarbiat Modares University, Iran
Natarajan Meghanathan	Jackson State University, USA
Nicolas Sklavos	Technological Educational Institute of Patras, Greece
Nidaa Abdual Muhsin Abbas	University of Babylon, Iraq
Olakanmi Oladayo	University of Ibadan, Nigeria
P. Ashok Babu	Narsimhareddy Engineering college, India
Patrick Seeling	University of Wisconsin - Stevens Point, USA
PESN Krishna Prasad	Aditya Engineering College, India
Phan Cong Vinh	London South Bank University, UK
Ponpit Wongthongtham	Curtin University of Technology, Australia
Premanand K. Kadbe	Vidya Pratishthan's College of Engineering, India
Rafael Timoteo	University of Brasilia - UnB, Brazil
Raja Kumar M.	Universiti Sains Malaysia
Rajagopal Palsonkennedy	Dr. MGR University, India
Rajarshi Roy	IIT- Kharagpur, India
Rajendra Akerkar	Technomathematics Research Foundation, India
Rajesh Kumar P.	The Best International, Australia
Rajeshwari Hegde	BMS College of Engineering, India
Rajeswari Balasubramaniam	Dr. MGR University, India
Rajkumar Kannan	Bishop Heber College, India
Rakhesh Singh Kshetrimayum	Indian Institute of Technology, Guwahati, India
Ramayah Thurasamy	Universiti Sains Malaysia, Malaysia
Ramayah Thurasamy	Universiti Sains Malaysia, Malaysia
Razvan Deaconescu	University Politehnica of Bucharest, Romania
Reza Ebrahimi Atani	University of Guilan Iran
Rohitha Goonatilake	Texas A&M International University, USA
S. Geetha	Anna University - Tiruchirappalli, India
S. Hariharan	B.S. Abdur Rahman University, India
Sagarmay	Deb Central Queensland University, Australia
Sajid Hussain	Acadia University, Canada
Salah M. Saleh Al-Majeed	University of Essex, United Kingdom
Salim Lahmiri	University of Québec at Montreal, Canada
Samarendra Nath Sur	Sikkim Manipal University, India
Sarmistha Neogy	Jadavpur University, India
Sattar B. Sadkhan	University of Babylon, Iraq
Sergio Ilarri	University of Zaragoza, Spain

Serguei A. Mokhov	Concordia University, Canada
Sharvani G.S.	RV College of Engineering, India
Shivan Haran	Arizona State University, USA
Shobha Shankar	Vidya Vardhaka College of Engineering, India
Shubhamoy Dey	Indian Institute of Management Indore, India
Sriman Narayana Iyengar	VIT University, India
Sundarapandian Vaidyanathan	VelTech Dr. RR & Dr. SR Technical University, India
SunYoung Han	Konkuk University, South Korea
Susana Sargento	University of Aveiro, Portugal
Virgil Dobrota	Technical University of Cluj-Napoca, Romani
Vishal Sharma	Metanoia Inc., USA
Wei Jie	University of Manchester, UK
Wichian Sittiprapaporn	Mahasarakham University, Thailand
William R. Simpson	Institute for Defense Analyses, USA
Xiaohong Yuan	North Carolina A & T State University, USA
Xin Bai	The City University of New York, USA
Yannick Le Moullec	Aalborg University, Denmark
Yaser M. Khamayseh	Jordan University of Science and Technology, Jordan
Yeong Deok Kim	Woosong University, South Korea
Yuh-Shyan Chen	National Taipei University, Taiwan
Yung-Fa Huang	Chaoyang University of Technology, Taiwan
Yung-Fa Huang	Chaoyang University of Technology, Taiwan
Zakaria Moudam	Université sidi mohammed ben Abdellah, Morocco
Nicolas Sklavos	Technological Educational Institute of Patras, Greece
Roberts Masillamani	Hindustan University, India

External Reviewers

Amit Choudhary	Maharaja Surajmal Institute, India
Abhishek Samanta	Jadavpur University, Kolkata, India
Anjan K.	MSRIT, India
Nana Patil	NIT Surat, Gujrat
Mydhili Nair	M.S. Ramaiah Institute of Technology, India
Padmalochan Bera	Indian Institute of Technology, Kharagpur, India
Osman B. Ghazali	Universiti Utara Malaysia, Malaysia
Suparna DasGupta	suparnadasguptait@gmail.com
Cauvery Giri	RVCE, India
Pradeepini Gera	Jawaharlal Nehru Technological University, India
Reshmi Maulik	University of Calcutta, India
Soumyabrata Saha	Guru Tegh Bahadur Institute of Technology, India
Srinivasulu Pamidi	V.R. Siddhartha Engineering College Vijayawada, India

Suhaidi B. Hassan	Office of the Assistant Vice Chancellor, Economics Building
Mahalinga V. Mandi	Dr. Ambedkar Institute of Technology, Bangalore, Karnataka, India
Omar Almomani	College of Arts and Sciences Universiti Utara Malaysia
Sara Najafzadeh	University Technology Malaysia
Ramin Karimi	University Technology Malaysia
Samodar Reddy	India School of Mines, India
Ashutosh Gupta	MJP Rohilkhand University, Bareilly
Jayeeta Chanda	jayeeta.chanda@gmail.com
Rituparna Chaki	rituchaki@gmail.com
Durga Toshniwal	Indian Institute of Techniology, India
Mohammad Mehdi Farhangia	Universiti Teknologi Malaysia (UTM), Malaysian
S. Bhaskaran	SASTRA University, India
Bhupendra Suman	IIT Roorkee (India)
Yedehalli Kumara Swamy	Dayanand Sagar College of Engineering, India
Swarup Mitra	Jadavpur University, Kolkata, India
R.M. Suresh	Mysore University
Nagaraj Aitha	I.T., Kamala Institute of Tech & Science, India
Ashutosh Dubey	NRI Institute of Science & Technology, Bhopal
Balakannan S.P.	Chonbuk Nat. Univ., Jeonju
Sami Ouali	ENSI, Compus of Manouba, Manouba, Tunisia
Seetha Maddala	CBIT, Hyderabad
Reena Dadhich	Govt. Engineering College Ajmer
Kota Sunitha	G. Narayanamma Institute of Technology and Science, Hyderabad
Parth Lakhiya	parth.lakhiya@einfochips.com
Murty, Ch. A.S.	JNTU, Hyderabad
Shriram Vasudevan	VIT University, India
Govardhan A.	JNTUH College of Engineering, India
Rabindranath Bera	Sikkim Manipal Inst. of Technol., India
Sanjay Singh	Manipal Institute of Technology, India
Subir Sarkar	Jadavpur University, India
Nagamanjula Prasad	Padmasri Institute of Technology, India
Rajesh Kumar Krishnan	Bannari Amman Inst. of Technol., India
Sarada Prasad Dakua	IIT-Bombay, India
Tsung Teng Chen	National Taipei Univ., Taiwan
Balaji Sriramulu	drsbalaji@gmail.com
Chandra Mohan	Bapatla Engineering College, India
Saleena Ameen	B.S. Abdur Rahman University, India
Babak Khosravifar	Concordia University, Canada
Hari Chavan	National Institute of Technology, Jamshedpur, India
Lavanya	Blekinge Institute of Technology, Sweden
Pappa Rajan	Anna University, India
Rituparna Chaki	West Bengal University of Technology, India

Salini P.	Pondichery Engineering College, India
Ramin Karimi	University Technology Malaysia
P. Sheik Abdul Khader	B.S. Abdur Rahman University, India
Rajashree Biradar	Ballari Institute of Technology and Management, India
Scsharma	IIT - Roorkee, India
Kaushik Chakraborty	Jadavpur University, India
Sunil Singh	Bharati Vidyapeeth's College of Engineering, India
Doreswamyh Hosahalli	Mangalore University, India
Debdatta Kandar	Sikkim Manipal University, India
Selvakumar Ramachandran	Blekinge Institute of Technology, Sweden
Naga Prasad Bandaru	PVP Siddartha Institute of Technology, India
HaMeEm sHaNaVaS	Vivekananda Institute of Technolgy, India
Gopalakrishnan Kaliaperumal	Anna University,chennai
Ankit	BITS, PILANI India
Aravind P.A.	Amrita School of Engineering India
Subhabrata Mukherjee	Jadavpur University, India
Valli Kumari Vatsavayi	AU College of Engineering, India

Contents

Signal, Image Processing and Pattern Recognition

Soft Computing, Artificial Intelligence

Algorithms and Applications

Automatic FAPs Determination and Expressions Synthesis

Narendra Patel and Mukesh A. Zaveri

Department of Computer Engg, BVM Engg. College, V.V. Nagar, India
Department of Computer Engg, SVNIT, Surat, Gujarat, India
bvm_nmp@yahoo.com, mazaveri@gmail.com

Abstract. This paper presents a novel method that automatically generates facial animation parameters (FAPs) as per MPEG 4 standard using a frontal face image. The proposed method extracts facial features like eye, eyebrow, mouth, nose etc. and these 2D features are used to evaluate facial parameters, namely called facial definition parameters using generic 3D face model. We determine FAPs by finding the difference between displacement of FDPs in specific expression face model and neutral face model. These FAPs are used to generate six basic expressions for any person with neutral face image. Novelty of our algorithm is that when expressions are mapped to another person it also captures expression detail such as wrinkle and creases. These FAPs can be used for expression recognition. We have tested and evaluated our proposed algorithm using standard database, namely, BU-3DFE.

Keywords: a generic 3D model, expression, texture, FAPs, FDPs, MPEG-4.

1 Introduction

The MPEG-4 visual standard specifies a set of facial definition parameters (FDPs) and FAPs for facial animation [1, 2]. The FDP defines the three dimensional location of 84 points on a neutral face known as feature points (FPs). The FAPs specify FPs displacements which model actual facial features movements in order to generate various expressions. The FAPs are a set of parameters defined in the MPEG-4 visual standard for the animation of synthetic face models. The FAP set includes 68 FAPs, 66 of which are low level parameters related to the movements of lips, jaw, eyes, mouth, cheek, nose etc. and the rest two are high-level parameters, related to visemes and expressions. All FAPs involving translation movement are expressed in terms of facial animation parameters unit (FAPU). FAP determination methods [3] are classified in two categories (1) Feature based and (2) optical flow based.

In feature based approach [3] areas containing the eyes, nose and mouth are identified and tracked from frame to frame. Approaches that are based on optical flow information [3, 5] utilize the entire image information for the parameter estimation leading to large number of point correspondences.

In order to precisely extract facial features, various approaches aimed at different sets of facial features have been proposed in a diversity of modalities.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 1–11.
springerlink.com

The mainstream approaches can be categorized into brightness-based and edge-based algorithms. Brightness-base algorithms exploit the brightness characteristics of the images to extract facial features. A typical approach of this class of algorithms is to employ the knowledge of the geometrical topology and the brightness characteristics of facial features, such as the eyebrows, eyes and mouth [6]. Edge-based algorithms target contours of the features, such as those of the mouth, eyes and chin, usually based on the gradient images. Hough transform, active contour model, i.e. snakes, and deformable templates are the edge based approaches used to detect facial features [7].

In this paper we have proposed a feature based approach for FAPs determination. We detect features like eye, eyebrow, nose and mouth from neutral face or expression specific frontal face using edge and brightness information. These features are used to adapt generic 3D face model into face specific 3D model. The displacement of FDPs in neutral 3D face model is determined as per MPEG 4 standard. Same way displacement of FDPs in different expressions specific 3D face model is determined. FAPs are determined by finding the difference between the displacement of FDPs in neutral 3D face model and specific expression 3D face model. Using these FAPS we have generated six basic expressions like anger, surprised, fear, sad, disgust and happy for any person whose neutral frontal face is available. The paper is organized as follows: Section 2 describes feature extraction. It is followed by FDP/FAP estimation and expressions generation in section 3 and 4 respectively. Section 5 describes expression mapping. The simulation results and conclusions are discussed in section 6 and 7 respectively.

2 Feature Extraction

Facial feature extraction comprises two phases: face detection and facial feature extraction. Face is detected by segmenting skin and non skin pixels. It is reported that YC_bC_r color model is more suitable for face detection than any other color model [8]. It is also reported that the chrominance component C_b and C_r of the skin tone always have values between $77<=C_b<=127$ and $133<= C_r <=173$ respectively [9]. After detection of face the features like eyes, mouth and eyebrows are detected.

2.1 Eye and Eyebrow Detection

After detection of face, the features like eyes, mouth and eyebrows are detected. We first build two separate eye maps, one from the chrominance components and the other from the luminance component [10].We have used upper half of the face region for preparation of eye maps to detect eyes. The eye map from the chroma is based on the observation that high C_b and low C_r values are found around the yes. It is constructed by

$$Ec = \frac{1}{3}\left((C_b^2) + (\bar{C}_r)^2 + \frac{C_b}{C_r} \right) \tag{1}$$

Where C_b^2, $(\bar{C}_r)^2$ and C_b / C_r all are normalized to the range [0 255] and $(\bar{C}_r)$ is the negative of C_r (i.e. 255-C_r).

The eyes usually contain both dark and bright pixels in the luma component so grayscale morphological operators dilation ($\oplus$) and erosion (Θ) is used to emphasis brighter and darker pixels in the luma component around eye regions. It is constructed using equation (2).

$$E_l = \frac{Y(x, y) \oplus G(x, y)}{Y(x, y) \Theta G(x, y)} \tag{2}$$

Where Y(x, y) is luma component of face region and g(x, y) is structuring element.

The eye map from the chroma is combined with the eye map from the luma by an AND (multiplication) operation. The resulting eye map is dilated with same structuring element to brighten eyes and suppress other facial areas. The locations of the eyes are estimated from the eye map. We have determined mean and standard deviation of eye map which is used to find location of eyes. After the large number of experiments we have set the value of threshold (T) =mean +0.3*variance. Eye feature points, the left and right corners and the upper and lower middle points of the eyelids are extracted from the edge map of the eye using sobel gradient operator. After two eye corners and two middle points on the eyelids have been located two parabolas are applied on the detected eyes. The location and feature points of the eyebrows are found from the edge map of the region of the face above the eye.

2.2 Lip Detection

Lip region is extracted using the observation that the lip pixels have stronger red component but green and blue components are almost same [11]. Skin pixels also have stronger red component but green component has higher value compared to blue component. Difference between red and green component is greater for lip pixels than skin pixels. Hulbert and poggio [12] proposed a pseudo hue definition that calculates pseudo hue as:

$$H(x, y) = \frac{R(x, y)}{R(x, y) + G(x, y)} \tag{3}$$

Where R(x, y) and G(x, y) are the red and green components of the pixel (x, y) respectively. However a person with reddish skin, pseudo hue may not give correct result. So we have combined pseudo hue H(x, y) with H1(x, y).

$$H1(x, y) = Log\left(\frac{G(x, y)}{B(x, y)}\right) \tag{4}$$

Where G(x, y) and B(x, y) are the green and blue components of the pixel (x, y) respectively. Lip pixels have lower value of Green and Blue color components so log function is used to enhance contrast. Lip pixels have higher value of H (x, y) and lower values of H1(x, y). The location of the mouth is detected by finding the region having higher value of H(x, y) and lower value of H1(x,y). We have found that pseudo hue (H) value varies from 0.55 to 0.65 and value of H1 is to be less than 0.73 for lip pixels. It is found that lip corners are in shadow and they have lower value of intensity. Lip corner points are found using intensity component of lip region having lower value.

The pseudo hue component H(x, y) of the lip region is shown in Figure 1. It is observed that the hue value (H) for the middle part of the lip pixels are higher when mouth is closed .But when moth is open the hue value is lower for teeth part but higher for cavity. This observation is used to check whether mouth is closed or open.

We have applied canny edge detector on intensity component of lip region and determined edge points corresponding to upper outer and lower outer lip contour for middle column. When mouth is closed, inner upper and inner lower boundary edge points are same. They are the points with maximum pseudo hue value for middle column as shown in Figure 1.

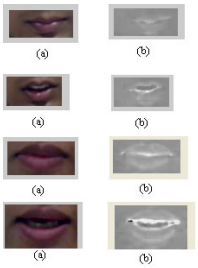


Fig. 1. (a) Lip region (b) Pseudo hue (H)

We have found P2, P3 and P4 points on the upper boundary of lip as shown in Figure 2. To find P2 we have traversed left edge of upper lip boundary from P4 till position is decreasing. P2 is an edge point with lowest position. Similarly we have traversed right edge of upper lip boundary to find point P3. When mouth is open feature points on inner upper lip boundary (P8) and inner lower lip boundary (P9) are determined. Teeth and tongue cause problems in determination of P8 and P9 from edge map.

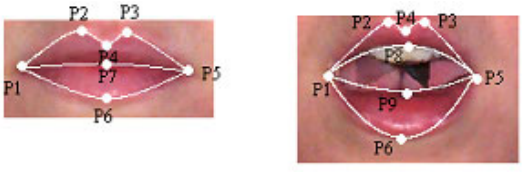


Fig. 2. Lip model

So after determination of P4 we have searched for first point in down direction up to P1 which has maximum gradient of pseudo hue (H) for middle column, which is P8. Same way after determination of P6 we have searched for first point in up direction up to P1 which has maximum gradient of pseudo hue (H) for middle column which is P9. After detecting feature points the upper lip boundary is modeled using cubic curve (cardinal spline) [13]. Experimentally it is found that upper inner boundary, lower inner boundary and lower outer boundary of lip can be modeled more accurately using parabola than cubic curve which is shown in Figure 2. The

location and feature point of nose is found using vertical component of gradient of the face image between eye and mouth.

3 FDP and FAP Generation

This is a process in which the generic 3D face model is deformed to fit a specific face [14]. Our proposed generic model [13] is shown in Figure 3 and Figure 4 which is polygon-based (triangle mesh) and consists of 350 triangles and 215 vertices. Model is adapted to given frontal face with the help of two geometrical transformations scaling and translation. Assuming orthographic projection, the translation vector can be derived by calculating the distance between the 3D face model centres to the 2D face centre. Let C_l indicate centre of left eye, C_r indicate centre of right eye, C_c indicate middle point between two eyes and C_m indicate centre of mouth in given face. Similarly C_l', C_r', C_c' and C_m' are corresponding points in the 2D projection of the 3D face model. Model is scaled by an amount S_x, S_y and S_z using equation (5)

$$S_X = \frac{|C_l - C_r|}{|C_l' - C_r'|} \qquad S_y = \left|\frac{C_c - C_m}{C_c' - C_m'}\right| \qquad S_z = \frac{S_x + S_y}{2} \tag{5}$$

After global adaptation of model we perform local refinement of model eyes, eyebrows, mouth and contour with that of face features. Appropriate translation factor does local refinement of the model. Constructed 3D models are shown in Figure 5.

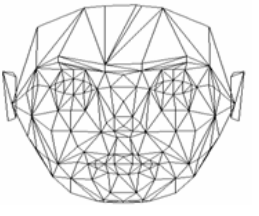


Fig. 3. Generic face model

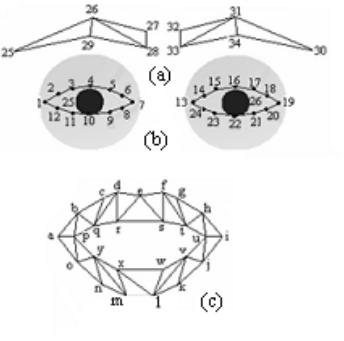


Fig. 4. Models of (a) Eyebrow (b) Eyes (c) Mouth

After we have modified generic face model, we can extract 3D coordinates of the FDP feature points from the model. We have determined position of FDPs corresponding to eye, eyebrow, and lip for neutral face image. We have also determined position of FDPs corresponding to eye, eyebrow, and lip for different facial expression models. We have determined FAP by finding difference between displacement of FDPs in specific expression face model and neutral face model. Measured FAPs for different expressions are shown in Table 1. They are measured by

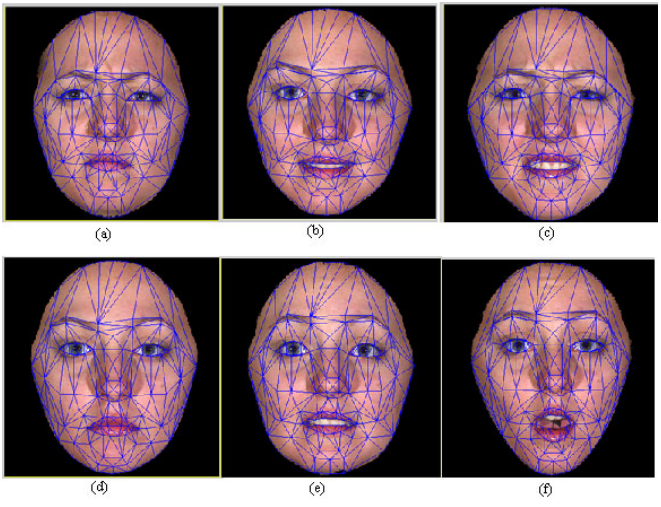


Fig. 5. constructed 3D models (a) angry (b) happy (c) disgust (d) sad (e) far (f) surprise

Table 1. Automatic determination of FAPs for different expressions

FAP	Happy	Sad	Disgust	Surprise	Anger	Fear
raise_l_i_eyebrow	1	-1	-5	29	-13	20
raise_l_o_eyebrow	0	-2	-17	8	-28	4
raise_l_m_eyebrow	0	-1	-14	18	-29	7
close_t_l_eyelid	1	3	8	-1	8	-3
lower_t_midlip_o	8	0	24	17	8	4
lower_t_midlip	10	-1	25	17	7	7
raise_b_midlip	-1	-1	7	-19	7	9
raise_b-midlip_o	-1	-1	13	-19	10	9
raise_l_cornerlip	17	-3	17	4	6	0
stretch_l_cornerlip	14	2	9	-13	2	1
squeeze_l_eyebrow	1	2	-4	0	-2	6

facial animation parameter units (FAPUs) that permit us to place FAPs on any facial model in a consistent way [15]. The FAPUs are defined with respect to the distances between key facial features in their neutral state such as eyes (ES0), eyelids (IRDS0), eye-nose (ENS0), mouth-nose (MNS0) and lip corners (MW0) as shown in Figure 6.

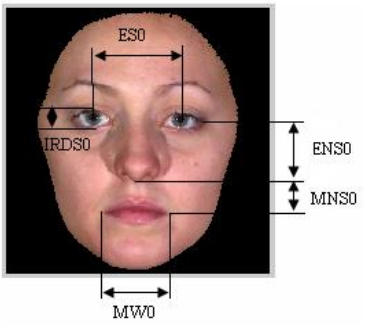


Fig. 6. Neutral face and FAPUs

4 Expressions Generation

Expressions are represented with the help of FAPs, a set of parameters defined in the MPEG-4 visual standard for the animation of synthetic face models. We have generated six basic expressions with the help of low level FAPS as discussed in [16]. The FAPS are computed through tracking a set of facial features and they are measured by facial animation parameter units (FAPUs) that permit us to place FAPs on any facial model in a consistent way [15]. The FAPUs are defined with respect to the distances between key facial features in their neutral state such as eyes (ES0), eyelids (IRDS0), eye-nose (ENS0), mouth-nose (MNS0) and lip corners (MW0) as shown in Fig. 6 and values are shown in Table 2.

Table 2. FAPUs

FAPUS	Value
ES	ES0/1024=0.1737
ENS	ENS0/1024=0.0902
MNS	MNS0/1024=0.0838
MW	MW0/1024=0.1133
IRISD	IRISD0/1024=0.0293

Table 3 gives the relation between expressions and involved FAPs. Expressions are generated by moving and deforming various control vertices of face model according to FAPs. If V_m indicates the neutral coordinate of the m^{th} vertex in a certain dimension of the 3D space, its animated position V_m' in the same dimension can be expressed as

$$V_m^{'} = V_m + w_n * FAPU_n * FAP_n \tag{6}$$

Where ω_n is the weight of the n^{th} FAP, $FAPU_n$ is the FAPU to n^{th} FAP and FAP_n is the amplitude of FAP. In fact, the term, ωn * FAPU*n* *FAPn defines the maximum displacement of m^{th} vertices. We have developed scan line algorithm [16] which establish correspondence between each triangle of neutral model and expression model for each scan line for each pixel to generate expression specific texture.

Table 3. Facial expressions and FAPs

Expressions	FAPs no
Happiness	Raise corner lip, stretch corner lip, mouth open 59,60,6,7,4,5,51,52
Sadness	Lower corner lip, lower inner eyebrow, close eyelid 59,60,31,32,19,20
Disgust	Close eyelid , mouth open , raise corner lip 19,20, 4,5,51,52,59,60
Surprise	Raise eyebrow , mouth open, open eyelid 31,32,33,34,35,36, 4,5,51,52,19,20
Anger	Open eyelid, lower eyebrow, squeeze eyebrow , mouth open 19,20 ,31,32,37,38, 4,5,51,52
Fear	eyebrow, mouth open 19,20 ,31,32,33,34,35,36,37,38, 4,5,51,52

5 Expression Mapping

Our proposed algorithm determines FAP from the given expressions database. It constructs 3D model from the frontal neutral face of any person and using these FAPs it also successfully map expressions of one person onto another person. To take care of the expression details such as wrinkle we have to also consider illumination changes along with geometry deformation. Let A and A' are the images of person A's neutral and expression face respectively. Let B denotes person B's neutral face. Our algorithm first determines FAPS from the A's expression image. Then the method discussed in section 4 is used to generate expression image of B denoted as B'. Now make it more realistic we also consider illumination changes in calculating intensity at every pixel.

According to phong illumination model intensity at a given point is given as

$$I_{a=}K_{d*I_p}(L_a \cdot N_a) \tag{7}$$

Where L is light source direction and N is a normal to surface and Kd is diffuse reflection coefficient and Ip is light source intensity. Intensity of deformed expression model is determined as

$$I_a^{'} = K_d * I_p(L_a^{'} \cdot N_a^{'}) \tag{8}$$

Where Ia and Ia' are intensity at every pixel of neutral face A and expression face A' respectively. With the help of expression synthesis algorithm we have already found out expression image of B denoted as B'.

$$I_b = K_d * I_p * (L_b \cdot N_{b)} \tag{9}$$

$$I_b^{'} = K_d * I_p * (L_b^{'} \cdot N_b^{'}) \tag{10}$$

Since human faces have approximately same geometrical shape so their surface normal at the corresponding positions are same

$$N_a = N_b and \quad N_a^{'} = N_b^{'} \tag{11}$$

Same way light source direction is also same at the corresponding positions.

$$L_a = L_b \quad and \quad L_a^{'} = L_b^{'} \tag{12}$$

So from above equations

$$\frac{I_a^{'}}{I_a} = \frac{I_b^{'}}{I_b} \Rightarrow \mathrm{I_b^{'}} = \frac{\mathrm{I_a^{'}}}{\mathrm{I_a}} * \mathrm{I_b} \tag{13}$$

I_a'/I_a is also known as expression ratio image (ERI) [17].After finding B' we also multiply intensity at every pixel with ratio I_a'/I_a which give more realistic image.

6 Simulation Results

We have used BU-3DFE standard database [18] which consists different person's neutral and expression frontal faces. The database covers both male and female images with different expressions, various nationality and different illuminations. We have evaluated our algorithm on many face images and result of 3D model construction are shown in Figure 5 which is used to locate FDPs. Table 2 shows the result of FAPs determined for all six basic expressions. Six basic expressions are generated using derived FAPs which are shown in Figure 7. The comparison between

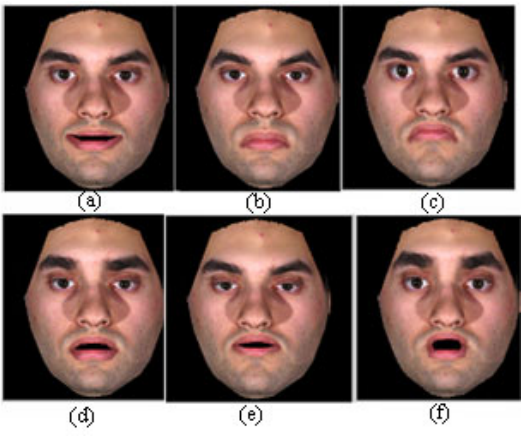


Fig. 7. Synthesize expressions (a) happy (b) sad (c) angry (d) fear (e) disgust (f) surprise

determination of FAPs with manually tracking facial features and with our algorithm is shown in Figure 8. Figure 9 (b) is synthesized angry expression with expression ratio method which looks more realistic compared to Figure 9 (c) considering only geometry deformation.

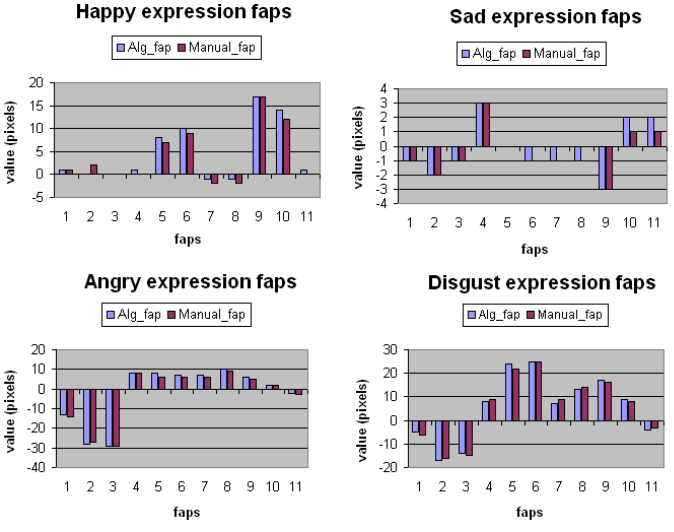


Fig. 8. Comparison among FAPs determination methods

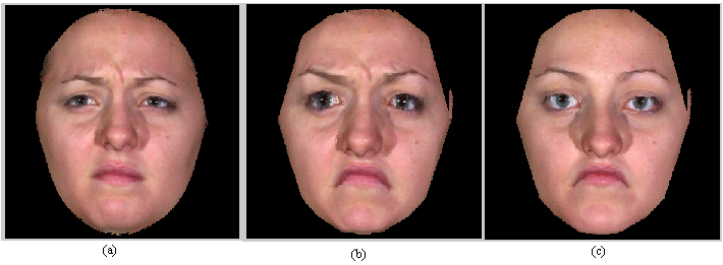


Fig. 9. (a) angry expression image (b) synthesized expression image with expression ratio image method (c) synthesized expression with geometry deformation method

7 Conclusion

In this paper we have proposed an automatic FAPs determination method using single frontal face image. Our proposed algorithm find features and feature points accurately from frontal face image with any expression. Results show that FAPs found using our proposed algorithm are very near to actual value. Using these FAPs we have successfully generated basic six expressions for any person whose frontal neutral face image is available. The expression mapping with expression ratio image is more realistic and expressive than traditional geometry deformation method.

Acknowledgment. The authors would like thank to state university of New York at Binghamton for providing database BU-3DFE.

References

[1] Krindis, S., Pitas, I.: Statistical analysis of human facial expressions. Journal of Information Hiding and Multimedia Signal Processing 1(3), 241–260 (2010)
[2] Pandzic, S., Komiya, R., Forchheimer, R.: MPEG-4 facial animation: the standard, implementation and applications. John Wiley and Sons (2002) ISBN: 0-470-84465-5
[3] Sarris, N., Strintzis, M.G.: 3D modeling and animation: synthesis and analysis Techniques for the human body, illustrated edition, March 22. IRM press (2005)
[4] Kim, J.W.: Automatic FDP/FAP generation from an image sequence. In: IEEE International Symposium on Circuits and Systems, May 28-31, vol. 1, pp. 40–43 (2000)
[5] Ravyse, I., Sahli, H.: Facial Analysis and Synthesis Scheme. In: Blanc-Talon, J., Philips, W., Popescu, D., Scheunders, P. (eds.) ACIVS 2006. LNCS, vol. 4179, pp. 810–820. Springer, Heidelberg (2006)
[6] Sheng, Y., Sadka, A.H., Kondoz, A.M.: An automatic algorithm for facial feature extraction in video applications. In: Proc. of 5th International Workshop on Image Analysis for Multimedia Interactive Services, lisbon, Portugal, April 21-23 (2004)
[7] Xu, C., Prince, J.: Snakes, shapes and gradient vector flow. IEEE Trans. on Image Processing 17(3), 359–369 (1998)
[8] Ip, H.H.S., Yin, L.: Constructing a 3D individualized head model from two orthogonal views. The Visual Computer 12(5), 254–266 (1996)
[9] Feng, G.C., Yuen, P.C.: Recognition of head and shoulder face image using virtual frontal view image. IEEE Trans. Systems, Man and Cybernetics, Part A 30(6), 871–882 (2000)
[10] Hsu, R.L., Abdel-Mottaleb, M., Jain, A.K.: Face detection in color image. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(5), 696–706 (2002)
[11] Goecke, R., Millar, J.B., Zelinsky, A., Robert-Ribes, J.: Automatic Extraction of Lip Feature Points. In: Proceedings of the Australian Conference on Robotics and Automation ACRA, Melbourne, Australia, August 30 -September 1, pp. 31–36 (2000)
[12] Hulbert, A., Poggio, T.: Synthesizing a colour algorithm from examples. Sciences 239, 482–485 (1998)
[13] Patel, N.M., Patel, P., Zaveri, M.: Parametric model based facial animation synthesis. In: International Conference on Emerging Trends in Computing, Kamraj College of Engg. & Tech., Tamilnadu, India, January 8-10 (2009)
[14] Sheng, Y., Sadka, A.H., Kondoz, A.M.: Automatic single view based 3D face synthesis for unsupervised multimedia applications. IEEE Transactions on Circuits and System for Video Technology 18(17), 961–974 (2008)
[15] Zhang, Y., Zhu, Z., Yi, B.: Dynamic facial expression analysis and synthesis with MPEG-4 facial animation parameters. IEEE Transactions and Systems for Video Technology 18(10), 1383–1396 (2008)
[16] Patel, N.M., Zaveri, M.: 3D Facial Model Reconstruction, Expressions Synthesis and Animation using single frontal face image. International Journal of Signal, Image and Video Processing (SIVP) (2011), doi:10.1007/s11760-011-0278-9
[17] Liu, Z., Shan, Y., Zhang, Z.: Expressive expression mapping with ratio images. In: SIGGRAPH, Los angles, August 12-17, pp. 271–276 (2001)
[18] Yin, L., Sun, X., Wang, Y., Rosato, M.J.: A 3D facial expression database for facial behavior research. In: Proc. of Int. Conf. on FGR, UK, April 2-6, pp. 211–216 (2006)

Generation of Orthogonal Discrete Frequency Coded Waveform Using Accelerated Particle Swarm Optimization Algorithm for MIMO Radar

B. Roja Reddy[1] and M. Uttara Kumari[2]

[1] Department of Telecommunication Engineering, R.V. College of Engineering, Bangalore, India
rojareddyb@rvce.edu.in
[2] Department of Electronics and Communication Engineering, R.V. College of Engineering, Bangalore, India
uttarakumari@rvce.edu.in

Abstract. Design of orthogonal code sets with correlation properties can effectively improve the radar performance by transmitting specially designed orthogonal Multiple Input Multiple Output (MIMO) radar. A novel particle swarm algorithm is proposed to numerically design orthogonal Discrete Frequency Waveforms and Modified Discrete Frequency Waveforms (DFCWs) with good correlation properties for MIMO radar. We employ Accelerated Particle Swarm Optimization algorithm (ACC_PSO), Particles of a swarm communicate good positions, velocity and accelerations to each other as well as dynamically adjust their own position, velocity and acceleration derived from the best of all particles. The simulation results show that the proposed algorithm is effective for the design of DFCWs signal used in MIMO radar.

Keywords: Multiple Input and Multiple Output (MIMO) Radar, Discrete Frequency Coded waveform (DFCW), Accelerated Particle Swarm Optimization Algorithm (ACC_PSO).

1 Introduction

Recently, the concept of MIMO radars has drawn considerable attention due to their ability to image a target from a variety of angles, thus improving the information received from the target and capability of modern radars in terms of detection, identification and classification of target under surveillance [1].

One of the important properties for MIMO radar is the resolution performance enhancement. The range resolution can be significantly improved by using very short pulses. This results in the decrease in received signal to noise ratio. To increase signal to noise ratio various Pulse compression techniques were developed. Although there are analog pulse compressions techniques, digital pulse compression is more popularly used. These include frequency codes, binary phase codes, and poly phase

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 13–23.
springerlink.com

codes. The performance criteria of a pulse compression sequence are judged by their autocorrelation properties [2].

The orthogonal waveforms used by the MIMO radar systems must be carefully designed to avoid the self-interference and detection confusion. For high range resolution and multiple target resolution, the aperiodic autocorrelation functions of sequences should have low of peak sidelobes level. Design of Orthogonal code sets with low Autocorrelation side lobe peaks (ASP) and cross correlation peaks (CP) is crucial for the implementing MIMO radar systems.

There are several types of waveforms to design orthogonal waveforms, such as polyphase waveform Discrete Frequency-coding Waveform (DFCW) which has large compressed ratio. The polyphase compression ratio is relatively small when compared with DFCW. The Autocorrelation Sidelobes Peak (ASP) level of discrete frequency-coding waveform with fixed frequency pulses is very large. By replacing fixed frequency pulse with linear Frequency Modulation Pulses can lower ASP but the grating lobes will appear in the range response if $T\Delta f > 1$ [2]. The relationship between subpulse duration T, frequency step Δf and LFM bandwidth B is set to make the grating lodes disappear [2].

Deng [3] has proposed simulated annealing algorithm to design of DFCW and got good results. Liu has proposed an approach using optimization Genetic Algorithm (GA) [2] technique and a Modified Genetic Algorithm (MGA) [4] technique to design multiple orthogonal discrete frequency-coding sequences with good aperiodic correlation.

In this paper, we employ ACC_PSO to design discrete frequency-coding waveform to obtain good correlation properties. To achieve this object the cost function was designed based on Peak Side lobe level Ratio and Integrated Side lobe Level Ratio. This algorithm as an optimization engine to obtain an optimal solution to this problem and also stabilizes to the solution in considerably lesser computational efforts. In this algorithm the Particles of a swarm communicate good positions to each other as well as dynamically adjust their own position, velocity and acceleration derived from the best position of all particles.

The rest of paper is organized as follows. In section 2 we formulate the waveform design using DFCW. In section 3 we introduce ACC_PSO to numerically optimize DFCW. Design results from proposed algorithm in section 4. Finally some conclusions are drawn in section 5.

2 The Waveform Design

Linear Frequency Modulation (LFM) the most popular pulse compression method. The basic idea is to sweep the frequency band B linearly during the pulse duration T. B is the total frequency deviation and the time bandwidth product of the signal is BT. The spectral efficiency of the LFM improves as the time-bandwidth product increases, because the spectral density approaches a rectangular shape. Here we consider the sequence length of each waveform (N) as 32 and Number of antennas (L) as 3.

2.1 DFCW with Frequency Hopping

The DFCW_FH waveform is defined as [4]

$$S_p(t) = \sum_{n=0}^{N-1} A(t) e^{j2\pi f_n^p t} \tag{1}$$

Where $A(t) = \begin{cases} 1/T \longrightarrow 0 \le t \le T \\ 0 \longrightarrow elsewhere \end{cases}$ and p= 1, 2, …, L T is the subpulse time duration. N is the number of subpulse that are continuous with the coefficient sequence $\{n_1, n_2, n_3, ...n_N\}$ with unique permutation of sequence $\{0.1,2,3,.....N-1\}$. f_n^p = n .Δf is the coding frequency of subpulse n of waveform p in the waveform DFCW_FH. Δf is the frequency step.

2.2 DFCW with Linear Frequency Modulation

By adding LFM to the DFCW_FH,the DFCW-LFM waveform is defined as [2]

$$S_p(t) = \sum_{n=0}^{N-1} e^{j2\pi f_n^p (t-nT)} . e^{j\pi k t^2}$$

$$\text{Where } A(t) = \begin{cases} 1/T \longrightarrow 0 \le t \le T \\ 0 \longrightarrow elsewhere \end{cases} \tag{2}$$

Where k is the frequency slope, $k = \frac{B}{T}$.

2.3 Modified DFCW with Linear Frequency Modulation

The modified DFCW waveform is proposed in this paper and is defined as

$$S_p(t) = \sum_{n=0}^{N-1} e^{j2\pi f_n^p (t-nT)} . e^{j\pi k t^3}$$

$$\text{Where } A(t) = \begin{cases} 1/T \longrightarrow 0 \le t \le T \\ 0 \longrightarrow elsewhere \end{cases} \tag{3}$$

The relationship between subpulse duration T, frequency step Δf and LFM bandwidth B is set to make the grating lobes disappear, it also lower the ASP.The B and T for each pulse remains a constant. The Δf values are called frequency steps. Each LFM pulse has a different starting frequency. The choice of BT, T.Δf and B/Δf values are crucial for the waveform design. Different lengths of firing sequence, N, have different values for each of the above mentioned parameters[2].

Table 1. Length of firing sequences & related parameters

Table .	B.T	B/Δf	T.Δf
8	18	6	3
16	36	12	3
32	72	24	3
64	144	48	3
128	288	96	3

The performance analysis is expressed in terms of Peak Side Lobe Ratio and integrated Sidelobe Level Ratio. The PSLR is a ratio of the peak sidelobe amplitude to the main lobe peak amplitude and is expressed in decibels. The autocorrelation and Cross correlation PSLR are defined as [7].

$$PSLR_{A} = 20\log_{10}\left\{\frac{\max_{n} \in sidelobe \left|A(S_l, n)\right|}{\max_{n} \in mainlobe \left|A(S_l, n)\right|}\right\}$$
$$PSLR_{C} = 20\log_{10}\left\{\frac{\max_{n} \in sidelobe \left|C(S_p, S_q, n)\right|}{\max_{n} \in mainlobe \left|C(S_p, S_q, n)\right|}\right\} whereq \neq p \tag{4}$$

The ISLR is a ratio of the integrated energy of all side lobes which spread across the whole time domain to the integrated energy of the main lobe in the pulse compression function.

$$ISLR_{S} = 20\log_{10}\left\{\frac{\sum onlySidelbe\left|A(S_p, n)\right|\sum\left|C(S_p, S_q, n)\right|}{\sum onlymainlbe\left|A(S_p, n)\right|}\right\} whereq \neq p \tag{5}$$

The cost function can be written as

$$CF = \sum_{l=1}^{L}\left(\min(PSLR_{Al}) + \min(PSLR_{Cl})\right) + \lambda \cdot \sum_{l=1}^{L} ISLR_{sl} \tag{6}$$

where λ is the relative weigh assigned to the ISLR and PSLR.

3 Accelerated Particle Swarm Optimization Algorithm

A lot of optimization methods have been developed for solving different types of optimization problems in recent years. There is no known single optimization method available for solving all optimization problems.

The modern optimization methods are very powerful and popular methods for solving complex engineering problems. These methods are particle swarm optimization algorithm, neural networks, genetic algorithms, artificial immune systems, and fuzzy optimization.

The particle Swarm concept originated as a simulation of simplified social systems. The original intent was to graphically simulate the graceful but unpredictable choreography of a bird flock. At some point in the evolution of the algorithm, it was realized that the conceptual model was, in fact, an optimizer. Through a process of trial and error, a number of parameters extraneous to optimization were eliminated from the algorithm, resulting in the simple original implementation.

The Particle Swarm Optimization algorithm is a novel population-based stochastic search algorithm and an alternative solution to the complex non-linear optimization problem. The PSO algorithm was first introduced by Dr. Kennedy and Dr. Eberhart in 1995 and its basic idea was originally inspired by simulation of the social behavior of animals such as bird flocking, fish schooling and so on. It is based on the natural process of group communication to share individual knowledge when a group of birds or insects search food or migrate and so forth in a searching space, although all birds or insects do not know where the best position is. But from the nature of the social behavior, if any member can find out a desirable path to go, the rest of the members will follow quickly [5][6].

As compared with other optimization methods, it is faster, cheaper and more efficient. In addition, there are few parameters to adjust in PSO. That's why PSO is an ideal optimization problem solver in optimization problems. PSO is well suited to solve the non-linear, non-convex, continuous, discrete, integer variable type problems.

In ACC_PSO, each member of the population is called a particle and the population is called a swarm. Starting with a randomly initialized population and moving in randomly chosen directions, each particle goes through the searching space and remembers the best previous positions, velocity and accelerations of itself and its neighbors. Particles of a swarm communicate good position, velocity and acceleration to each other as well as dynamically adjust their own position, velocity and acceleration derived from the best position of all particles. The next step begins when all particles have been moved. Finally, all particles tend to fly towards better and better positions over the searching process until the swarm move to close to an optimum of the fitness function as shown in fig1.

ACC_PSO has been used as a robust method to solve optimization problems in a wide variety of applications. In ACC_PSO for the optimization we have considered

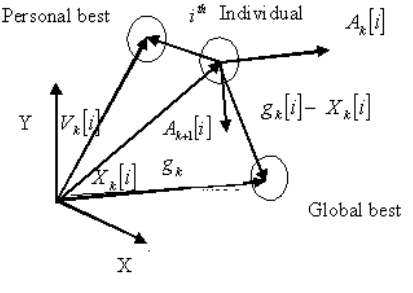


Fig. 1. Behavior of a individual in 2–dimensional search

three parameters position, velocity and acceleration for each swarm particle, where as in PSO only two parameters position and velocity are considered for each particle. Here in this algorithm the swarms are the random sequence and rand positions are generated. From these positions, the velocity and acceleration are generated.

The following steps are followed to optimize the sequence.

1. Generate the individuals $x_o[i], i \in [0,2,....N-1]$ of initial generation (k=0) randomly.
2. For each particle, compute the position, velocity and acceleration and update $p_x[i], V_x[i], A_x[i]$ for all individuals.
3. Compute the level for each particle depending on the weigh of each particle. The new vector generated is considered as swarm particle.
4. Compute the best acceleration for each particle and the best acceleration for all particles and called as local best and global beat in the iterations.
5. Update the new acceleration and add it to the swarm particle and get the new particle.

$$x_{k+1}[i] = x_k[i] + A_k[i], \forall\ i$$
$$A_k[i] = K.\{A_{k-1}[i] + c_1.\Delta(g - x_k[i]) + c_2.\omega(p_k(i) - x_k[i])\}, \forall i \tag{7}$$

Where, $c_1, c_2 and K \in R$ are weighting coefficients.

$$\Delta = diag[\alpha_1, \alpha_2, ...\alpha_d]\ \ \omega = diag[\beta_1, \beta_2, ...\beta_d]$$

Where $\alpha_i \in [0,1], \beta_i \in [0,1]$

where i is an pseudorandom numbers.

6. After updating all the particles convert the level vector to the weigh of each particle and compute Cost Function mentioned in equation (6). If the fitness function is satisfied the process ends other wise the whole process is repeated from step 3.

4 Design Results

Based on the proposed algorithm in section 3, the Discrete Frequency Coded Waveform (DFCW) set with the length of 32 and code of 3 are designed. In this paper the autocorrelation and cross correlation are normalized with respect to sequence length. The Optimized sequences of DFCW_FH, DFCW_LFM and Modified DFCW_LFM waveforms are generated using ACC_PSO. The simulation is carried out in Matlab. DFCW pulses have different starting frequencies with different combinations of BT, T.Δf and B/Δf. The Table1 shows the relation between the various parameters. The unique permutation of sequence generated by ACC_PSO is tabulated in the Table2. From Table3 to 5 show the ASPs and CPs for DFCW_FH, DFCW_LFM and Modified DFCW_LFM for the sequences shown in Table2. From the tabulated results it can be observed that the results ASPs and CPs of modified DFCW_LFM are better than that of DFCW_LFM and DFCW_FH.

The Autocorrelation and cross correlation functions for the 3 waveforms of DFCW_LFM for the sequence in Table 2 are shown in Fig2 and Fig 3 respectively.

The effect of Doppler on the designed DFCW_LFM waveform is considered. The output of the matched filter of the designed DFCW_LFM is reduced by considering $f_d T = 0.031$ [2]. The Partial ambiguity function for one designed waveform is plotted in the Fig 4 and the complete ambiguity function for the designed waveform is plotted in Fig 5. Convergence of cost function for Genetic Algorithm and Particle Swarm Optimization algorithm and ACC_PSO is plotted in Fig 6. The convergence of ACC_PSO is faster.

ACC_PSO algorithm is compared with the existing Simulated Annealing algorithm & Genetic Algorithm and Modified Genetic Algorithm in Table 6. The results show an improvement in ASPs and CPs. It infers that sequences generated by ACC_PSO algorithm have good correlation properties and also converges faster. The result of ACC_PSO with DFCW_LFM and Modified DFCW_LFM in Table 6 shows that auto correlation in Modified DFCW_LFM is improved over DFCW_LFM.

Table 2. The 3 sequences generated using ACC_PSO for DFCW

Table.	seq1	seq2	Seq3	Sl no.	seq1	seq2	Seq3
1	19	16	25	17	3	0	6
2	23	26	15	18	24	6	2
3	4	8	9	19	6	5	8
4	27	28	18	20	25	20	1
5	14	25	13	21	1	24	11
6	11	4	0	22	31	19	30
7	8	14	20	23	9	2	5
8	16	11	22	24	29	12	19
9	0	23	3	25	18	21	4
10	21	10	28	26	22	9	14
11	15	13	24	27	12	1	16
12	30	7	7	28	7	15	27
13	10	30	12	29	20	27	29
14	2	3	17	30	17	31	10
15	13	29	21	31	28	18	31
16	26	22	23	31	5	17	26

Table 3. ASPs and CPs of the designed DFCW_FH waveform using ACC_PSO for N=32 L=3

Table .	Code1	Code2	Code3
Code1	0.1316	0.0955	0.0711
Code2	0.0955	0.1045	0.0695
Code3	0.0711	0.0695	0.1154

Table 4. ASPs and CPs of the designed DFCW_LFM waveform using ACC_PSO for N=32, L=3

Table .	Code1	Code2	Code3
Code1	0.07957	0.0684	0.0468
Code2	0.0684	0.07048	0.049
Code3	0.0468	0.049	0.05066

Table 5. ASPs and CPs of the designed Modified DFCW_LFM waveform using ACC_PSO for N=32, L=3

Table .	Code1	Code2	Code3
Code1	0.09802	0.0622	0.0521
Code2	0.0622	0.04612	0.0482
Code3	0.0521	0.0482	0.0315

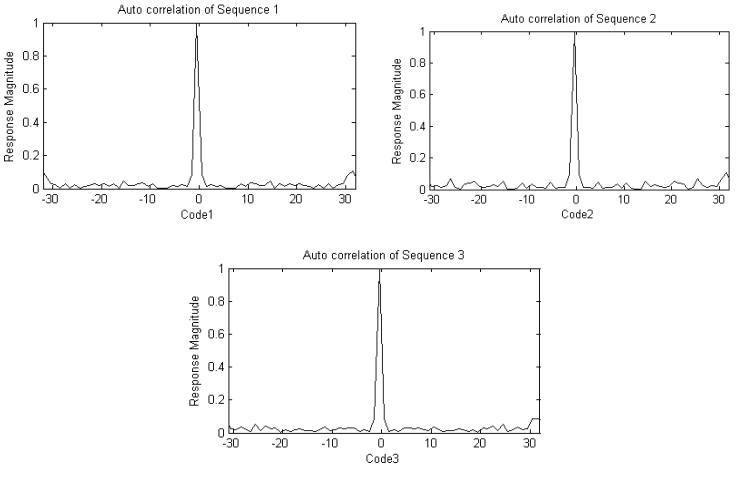


Fig. 2. Auto Correlation functions of sequence length N=32 and Antenna=3

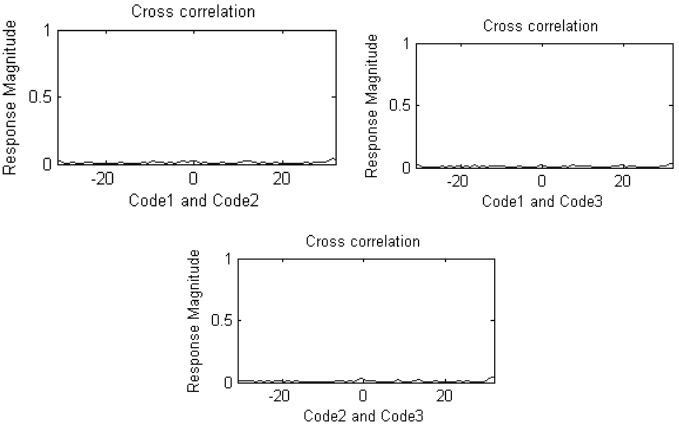


Fig. 3. Cross Correlation functions of sequence length N=32 and Antenna=3

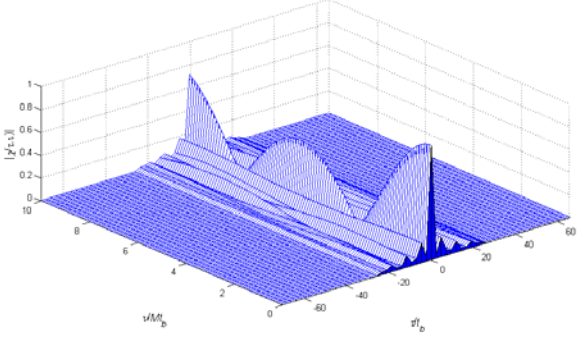


Fig. 4. Partial Ambiguity function for sequence length N=32 at one receiver

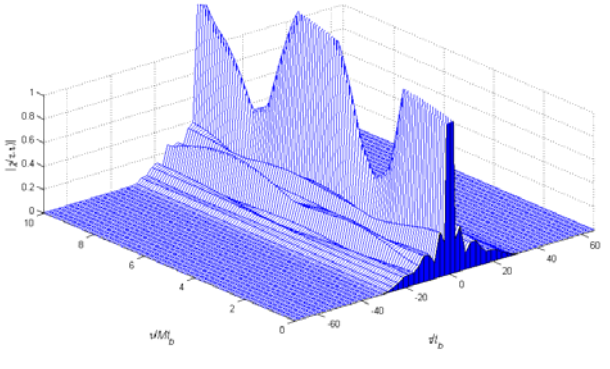


Fig. 5. Total Ambiguity function for sequence length N=32 at one receiver

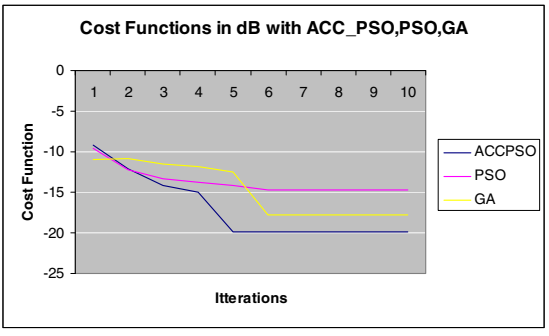


Fig. 6. Convergence of Cost Function in dB

Table 6. ASPs and CPs for Various Algorithms in Literature

Table .	ASPs result in dB	CPs result in dB
Simulated Annealing algorithm[3]	-21.5	-19.18
Genetic Algorithm DFCW_FH[4]	-13.9	-23.25
Genetic Algorithm DFCW_LFM[2]	-23.81	-24.94
ACC_PSO DFCW_FH	-18.62	-22.08
ACC_PSO DFCW_LFM	-23.49	-25.23
ACC_PSO Modified DFCW_LFM	-24.6	-25.3

5 Conclusions

In this paper, we have presented a new numerically optimized method of discrete frequency-coding waveform for orthogonal MIMO radar. This method is applicable to the case where the transmitted waveforms are orthogonal. The proposed method applies the Accelerated particle swarm optimization algorithm can get superior correlation properties to any existing sequences in literature. This approach is an alternative tool for the design of multiple orthogonal discrete frequency coding sequences with good correlation. The effectiveness of the proposed algorithm was demonstrated through the design results and compared with the literature values.

References

[1] Fishler, E., Haimovich, A., Blum, R., Cimini, L., Chizhik, D., Valenzuela, R.: MIMO radar: An idea whose time has come. In: Proceedings of IEEE International Radar Conference, Philadelphia, PA (April 2004)
[2] Liu, B., He, Z., Li, J.: Mitigation of Autocorrelation sidelobe peaks of Orthogonal Discrete Frequency-Coding waveform for MIMO Radar. In: Proceedings of IEEE Radar Conference, China, Chengdu, pp. 1–6 (2008)
[3] Deng, H.: Discrete Frequency-Coding Waveform Design for netted Radar Systems. IEEE Signal Processing Letters 11(2), 179–182 (2004)
[4] Liu, B., He, Z.: Orthogonal Discrete Frequency-Coding waveform for MIMO Radar. Spinger Link Journal of Electronics (China) 25(4) (July 2008)
[5] Kobayashi, T., Nakagawa, K., Imae, J., Zhai, G.: Real Time Object tracking on Video Image Sequence using Particle Swarm Optimization. In: International Conference on Control, Automation and Systems, Seoul, Korea, pp. 1773–1778 (2007)
[6] Deng, Y., Tong, H.: Dynamic Shortest path in stochastic Traffic networks baed on fluid neural network and particle swarm optimization. In: Internal Conference on Natural Computation (ICNC), Yantai Shandong, pp. 2325–2329 (2010)
[7] Bo, Z., Zhen, D., Nong, L.D.: Design and performance analysis of Orthogonal coding design in MIMO SAR. The Journal Science China, Information Science 54(8), 1723–1737 (2011)

Authors

B. Roja Reddy received the B.E degree in 1998 from Gulbarga University, Karnataka and M.E degree in 2004 from VTU, Karnataka. Presently working at R.V. college of Engineering with an experience of 9 years in the teaching field. Her research interest lies in various areas signal Processing. Currently précising her Ph.D in Radar Signal Processing.

M Uttara Kumari received the B.E degree in 1989 from Nagarujna University, Hyderabad, Andhra Pradesh and M.E degree in 1996 from Bangalore University, Karnataka and Ph.D degree in 2007 from Andhra University. Presently working at R.V. college of Engineering with an experience of 17 years in the teaching field. Her research interest lies in various areas of radar systems, Space-time adaptive processing, speech processing and image processing**.**

Text Independent Speaker Recognition Model Based on Gamma Distribution Using Delta, Shifted Delta Cepstrals

K. Suri Babu[1], Srinivas Yarramalle[2], and Suresh Varma Penumatsa[3]

[1] Scientist, NSTL (DRDO), Govt. of India, Visakhapatnam, India
[2] Dept. of IT, GITAM University, Visakhapatnam. India
[3] Aadikavi Nannaya University, Rajahmundry, India
suribabukorada2000@gmail.com, sriteja.y@gmail.com

Abstract. In this paper, we present an efficient speaker identification system based on generalized gamma distribution. This system comprises of three basic operations, namely speech features classification and metrics for evaluation. The features extracted using MFCC are passed to shifted delta cepstral coefficients (SDC) and then applied to linear predictive coefficients (LPC) to have effective recognition. To demonstrate our method, a database is generated with 200 speakers for training and around 50 speech samples for testing. Above 90% accuracy reported.

Keywords: Speaker identification, MFCC, LPC, Generalized Gamma, Shifted Delta coefficients.

1 Introduction

With the recent advancements in Technology, lot of information can be stored in the databases, in any of the format such as audio, video or text. Therefore, searching the exact information is difficult task [1]. Automatic indexing to the multimedia content can solve this problem. To retrieve speech signal from this Meta data is acrucial task.

The speech signal to be retrieved is considered and is divided in to small streams (segments) and the features are to be extracted.In order to extract features, MFCC are mostly proffered [3], [4] since they are less vulnerable to noise and give less variability. In order to have effective recognition it is needed to extract the first and second order time derivatives of cepstral features, that is delta and delta-delta features[5],but these features will be effective for short term speech samples, for long term features shifted delta coefficients (SDC) are well proffered[6], [7], [8].

Hence in this paper, we develop a model for speaker identification, where the features obtained from MFCC are converted to shifted delta coefficients and also by converting MFCC to delta coefficients. It is observed that the features obtained from MFCC followed by SDC outperform MFCC followed by delta.

The paper is organized as follows, the section-2 of the paper discuses about feature extraction, in section-3 generalized gamma distribution is proposed. Section -4 deals with experimental results. Finally, in section-5 conclusions are presented.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 25–29.
springerlink.com

2 Feature Extraction

In order to have an effective speaker identification model, the basic requirement is identifying the features effectively,in order to model the features MFCC are used along with the first order derivatives(delta coefficients) and second order derivatives (delta-delta coefficients),these combinations works effectively only for short duration speech signals and for longer duration speeches it is essential to use shifted delta cepstral coefficients along with MFCC for effective recognition since SDC reflects the dynamic cepstral features along with pseudo-Prosodic feature behavior[5].Hence this paper demonstrates the effectiveness of the usage by considering a database of 200 speakers for training and 50 speech samples for testing.

3 Generalized Gamma Mixture Model

Today most of the research in speech processing is carried out by using Gaussian mixture model, but the main disadvantage with Gaussian mixture model is that it relies exclusively on the approximation and low in convergence, and also if Gaussian mixture model is used, the speech and the noise coefficients differ in magnitude [7]. To have a more accurate feature extraction, maximum posterior estimation models are to be considered [8]. Hence in this paper,a generalized gamma distribution is utilized for classifying the speech signal. Generalized gamma distribution represents the sum of n-exponential distributed random variables both the shape and scale parameters have non-negative integer values [9]. Generalized gamma distribution is defined in terms of scale and shape parameters [10]. The generalized gamma mixture is given by

$$f(x,k,c,a,b)=\frac{c(x-a)^{ck-1}e^{-\left(\frac{x-a}{b}\right)^{c}}}{b^{ck}\Gamma(k)} \tag{1}$$

Where, k and c are the shape parameters, a is the location parameter, b is the scale parameter and gamma is the complete gamma function[11]. The shape and scale parameter of the generalized gamma distribution helps to classify the speech signal and identify the speaker accurately.

4 Experimental Results

During the training phase, the signal must be preprocessed and the features are extracted using MFCC. In order to have an effective recognition system we have sampled the data into short speech samples of different time frames and the MFCC features that are extracted are converted delta coefficients and shift delta coefficients.It is observed that MFCC combined delta coefficients could not effectively recognize the speech samples as compared to that of MFCC combined with SDC. The output is then fed to LPC (linear predictive coefficients). The features extracted are then given as input to the classifier that is generalized gamma distribution, using these feature set, the generalized gamma distribution is effectively recognized. The speech samples that are obtained from MFCC-SDC-LPC,it can also be seen that as and when the sample size is increased, these features that are extracted helps to classify the speakers most effectively. The results are presented in both tabular and graphical formats.

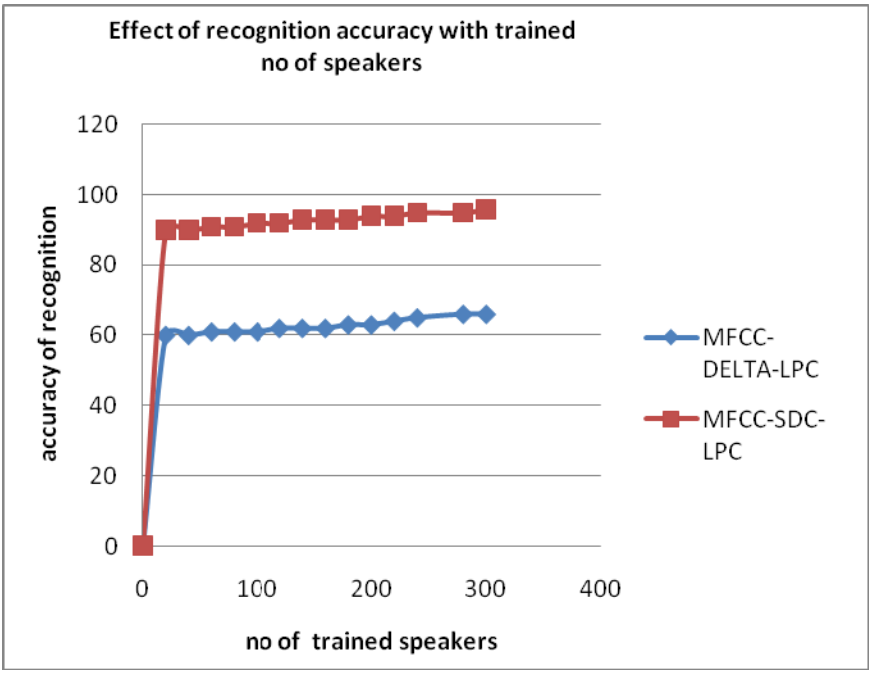


Fig. 1. Effect of Recognition accuracy with trained dataset

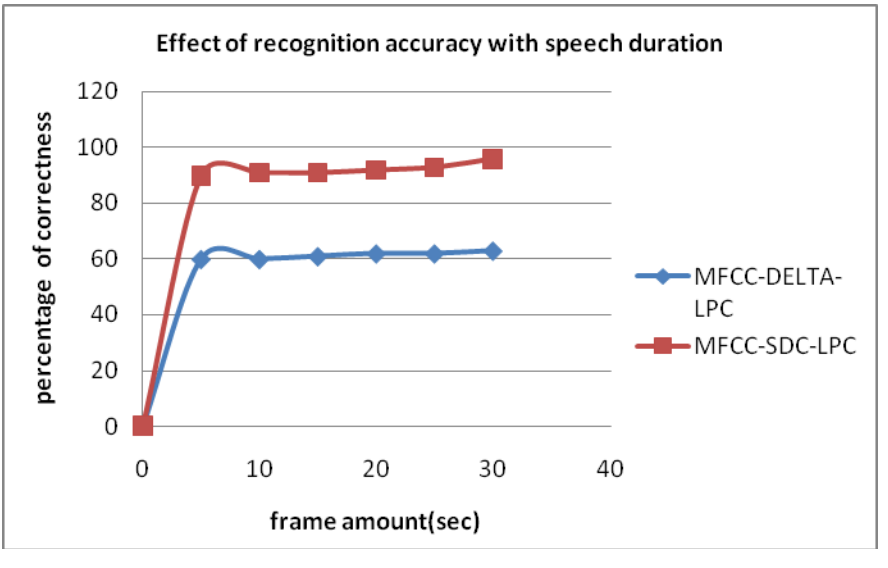


Fig. 2. Effect of Recognition accuracy with Speech duration

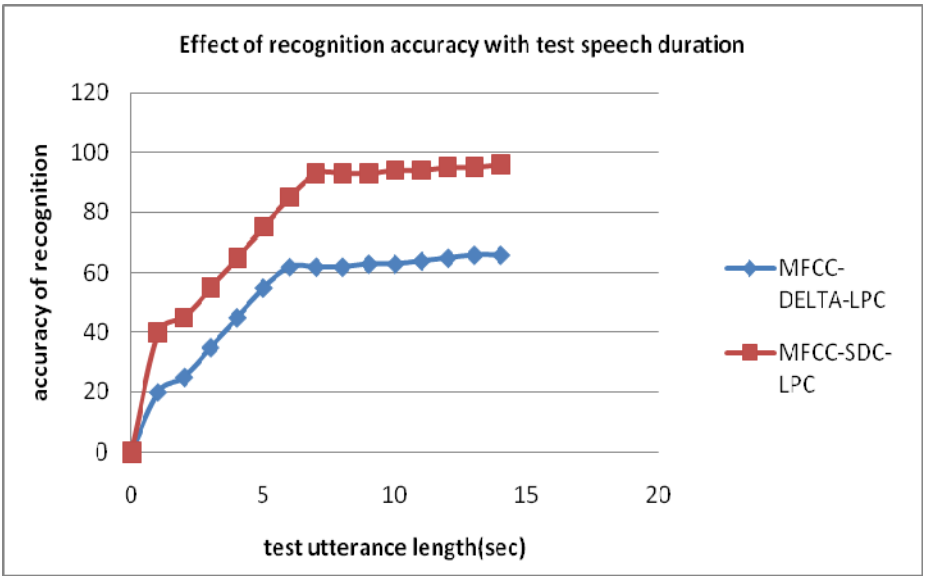


Fig. 3. Effect of recognition accuracy with test Speech duration

Table 1. Statistical data showing accuracy %

	No of trained speakers	Frame amount (sec)	Test utterance length (sec)	Recognition Accuracy (%)
MFCC-DELTA-LPC	0 to 50	0 to 5	0 to 5	Less than 60
	50 to 100	5 to 10	5 to 10	Around 60
	100 to 300	10 to 30	10 to 15	Above 62
MFCC-SDC-LPC	0 to 50	0 to 5	0 to 5	Less Than 80
	50 to 100	5 to 10	5 to 10	Around 85
	100 to 300	10 to 30	10 to 15	Above 90

From the above figures and table (Fig.1 to Fig.3 and Table 1),it could be easily seen that the MFCC-SDC-LPC outperforms MFCC-Delta-LPC and over all recognition rate is above 90% is seen in the developed model.

5 Conclusions

In this paper, we have developed a new model for speaker identification based on generalized gamma distribution. The speeches are extracted using MFCC are combined

with delta coefficients followed by LPC and also MFCC combined with SDC followed by LPC. The model is demonstrated a database of 200 samples and tested with 50 samples,the accuracy is around 90% and proved to be efficient model.

References

1. Kos, M., Vlaj, D., Kacic, Z.: Speaker's gender classification and segmentation using spectral and cepstral feature averaging. In: 18th International Conference on Systems, Signals and Image Processing, IWSSIP 2011 (2011)
2. Razik, J., Sénac, C., Fohr, D., Mella, O., Parlangeau-Valles, N.: Comparision of two speech/Music segmentation systems for audio indexing on Web. In: Proc. of WMSCI 2003, Florida, USA (July 2003)
3. Corneliu Octavian, D., Gavat, I.: Feature Extraction Modeling &Training Strategies in continuous speech Recognition For Roman Language. In: EU Proceedings of IEEE Xplore, EUROCN 2005, pp. 1424–1428 (2005)
4. Agarwal, S., et al.: Prosodic Feature Based Text-Dependent Speaker Recognition Using machine Learning Algorithm. International Journal of Engg. Sc. &Technology 2(10), 5150–5157 (2010)
5. Gonzalez, D.R., Calvo de Lara, J.R.: Speaker verification with shifted delta cepstral features: Its Pseudo-Prosodic Behaviour. In: Proc. I Iberian SLTech. (2009)
6. Torres-Carrasquillo, P.A., Singer, E., Kohlerand, M.A., Greene, R.J., Reynolds, A., Deller Jr., J.R.: Approches to language Identification Using Gausian Mixture Models and Shifted delta cepstral features. In: Proc. of ICSLP 2002, pp. 89–92 (2002)
7. Kinnunen, T., Koh, C.W.E., Wang, L., Li, H., Chang, E.S.: Temporal discrete cosine transform: Towards longer term temporal features for speaker verification. In: Proc of ICSLP 2006 (2006)
8. Calvo, J.R., Fernández, R., Hernández, G.: Channel / Handset Mismatch Evaluation in a Biometric Speaker Verification Using Shifted Delta Cepstral Features. In: Rueda, L., Mery, D., Kittler, J. (eds.) CIARP 2007. LNCS, vol. 4756, pp. 96–105. Springer, Heidelberg (2007)

Skin Segmentation Based Elastic Bunch Graph Matching for Efficient Multiple Face Recognition

Sayantan Sarkar

Department of Electrical Engineering, NIT Rourkela
sayantansarkar24@gmail.com

Abstract. This paper is aimed at developing and combining different algorithms for face detection and face recognition to generate an efficient mechanism that can detect and recognize the facial regions of input image.

For the detection of face from complex region, skin segmentation isolates the face-like regions in a complex image and following operations of morphology and template matching rejects false matches to extract facial region.

For the recognition of the face, the image database is now converted into a database of facial segments. Hence, implementing the technique of Elastic Bunch Graph matching (EBGM) after skin segmentation generates Face Bunch Graphs that acutely represents the features of an individual face enhances the quality of the training set. This increases the matching probability significantly.

Keywords: Elastic Bunch Graph Matching, Skin Segmentation, Gabor Wavelets, Graph Similarity, Template Matching, Face Bunch Graph, Face Recognition Database of University of Essex.

1 Introduction

Humans inherently use faces to recognize individuals and now, advancements in computing capabilities over the past few decades enable similar recognitions automatically [5]. Face recognition algorithms have, over the years, developed from simple geometric models to complex mathematical representations and sophisticated vector matching processes. [15]

Today, face recognition is actively being used to minimize passport fraud, support law enforcement agencies, identify missing children and combat identity theft.

Typically, the algorithms for face detection and recognition fall under any one of the following broad categories [19] –

1. **Knowledge-based methods:** encode what makes a typical face, e.g., the association between facial features.
2. **Feature-invariant approaches:** aim to find structure features of a face that do not change even when pose, viewpoint or lighting conditions vary(PCA).[17]
3. **Template matching:** comparison with several stored standard patterns to describe the face as a whole or the facial features separately.
4. **Appearance-based methods:** the models are learned from a set of training images that capture the representative variability of faces.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 31–40.
springerlink.com

2 Approach

As all the algorithms for facial recognition is based on certain set of specified features or template matching, a raw image input leads to over identification of features from the background region.

This over identified features leads to garbage values that alter the matching criteria leading to under identification or false identification.

To minimize such errors, a two-step approach is adopted by us that initially in the first step detects the facial region as the foreground and rejects the rest of the image segment as background. In the second step the facial recognition algorithm is executed only for the foreground region. Hence, the over identification is curtailed, increasing the efficiency of the algorithm with higher ratio of correct identification.

3 Image Enhancement

3.1 Contrast Stretching

The variance of the image is calculated and checked against a cut threshold, such that the non-uniform light regions get corrected. The contrast enhancement of the image may be done globally or adaptively. The contrast stretching function used is a simple piecewise linear function, although sigmoid functions may also be used.

3.2 Color Space Conversion

The standard input images are generally in RGB color space, which is also known as the visual color space. But in this Cartesian representation of the color space, the chroma and the luma components for each pixel is non-distinguishable. Hence, the color space is for each pixel in the image is converted to another popular color space (HSV color space).

4 Skin Segmentation

As the initial RGB image is already converted into HSV space after proper contrast stretching to remove luminous variations and dependence we can only focus on the hue and the saturation component [10]. These ‘H’ and ‘S’ color component are plotted, and averaged to give two final histograms. The histograms shows that the ‘H’ and ‘S’ components for faces are properly clustered.

In our analysis the highest point of the histograms were considered the Mean Value with a Variance of 50%. The suitable Gaussian Curve corresponding to the extracted data is then superimposed on the original histograms to threshold the image into two segments. The intra Gaussian segment corresponds to the skin segment.

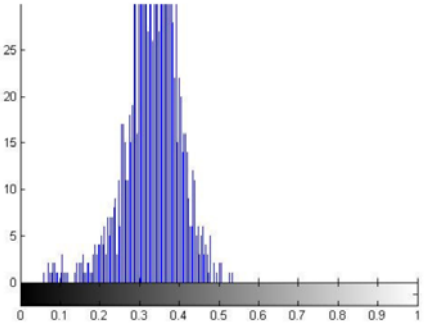


Fig. 1. Example Histogram of 'S' component of Training Facial image

4.1 Morphological Noise Removal

After the rejection of the non-skin region, the image still remains noisy due to stray pixels having skin chroma or non-facial skin regions that need to be cleared up [6]. The sequence of steps is further described as follows:

1. As the noise removal is intensity-based, the skin segmented image is converted to gray scale and threshold at a low threshold for background.
2. Morphological opening is performed using a small window convolution to remove the tiny skin segments, not classified as facial region and repeated with 17X17 window after filling intra facial holes.

0	1	0
1	1	1
0	1	0

Fig. 2. Structuring Window

4.2 Facial Region Classification

An image can contain skin region apart from the facial skin region. The skin segmented image is threshold on the basis of skin chroma, which makes the facial components such as eyes and nose to be threshold in the background as they do not match the skin chroma. They show up as 'holes' after proper binarizing via thresholding. The Euler numbers [14] for these binarized regions are then calculated.

An adaptive system is used to produce the threshold for each connected area. If there is a large spread and the ratio of mean to standard deviation is high, the threshold is set to a fraction of the mean. This stops darker faces from splitting up into many connected regions after thresholding.

After computation of the Euler number e. If $e < 0$ (i.e. less than two holes) the region is rejected. This is based on the fact that the face has at least two holes due to the presence of the eyes.

4.3 Template Matching

The Euler Number criteria are not limiting to facial segments. To remove this limitation average template is constructed manually. This template was convolved repeatedly with the image and the maximum peaks were extracted until the peak failed to exceed a calculated threshold value. The coordinates of the center of the bounding boxes which matched the most with this average template were stored in an array.

5 Detection

The primary motive of a good facial recognition algorithm is to correlate between a facial segment and a pre-identified facial database. But the performance criteria predominately depend on three factors: 1> Robustness 2> Speed 3> Hit Rate.

In our method the robustness and the speed of the algorithm is improved using the first step of the two step process, which thresholds the irrelevant back ground data. The input to the recognition algorithm is the facial segments extracted from the input images, which reduces the chances of false detection due to excess data points.

5.1 Elastic Bunch Graph Matching (EBGM)

The main idea of the EBGM is to design an algorithm that can comparatively analyze two images and find a quantitative value for the similarity between them [7].

The facial image in each bounding box is assumed to be threshold by skin segmentation such that only the skin region of the facial region is identified.

In our current analysis of the algorithm, we are assuming that the matching system is to be used as an integration of user identification systems, where due to the detection algorithm the facial segment is extracted and normalized with reference to a standard size and standard contrast. Hence, we assume that the distortions due to Position and Size are not present in out facial segment database.

So the only variation/ distortion are:

1. **Expression Distortions**: The facial expression of the person in the facial image is prone to change in every image hence cannot match the base database image.
2. **Pose Distortions**: The position of the person with respect to the camera can vary a little in the 2-D plane of the image but not over a large range.

The basic object representation used in case of EBGM is a graph. This graph is used to represent a particular face segment, is therefore a ‘facial graph’.

The facial graph is labeled with nodes and edges. In turn the nodes are identified with wavelet responses of local jets and the edges are identified by the length of the edge [3].

1. **Node:** It is a point in the 2-D facial graph of the image that signifies the phase and the magnitude of the wavelet response of the facial image in the locality of the node point.
2. **Edge**: It is a line that connects the nodes. Every two node in the graph is interconnected with an edge, which is represented with the magnitude of the length of the edge.

5.2 Create Bunch Graphs

Gabor Wavelets

For detection referring the earlier work of Wiscott[18]. Gabor wavelets generated using Gabor Wavelet Transform is applied as Gabor filters to wavelet space [11]. They are predominately edge detection filter that assigns magnitude and phase depending on edge directions and intensity in varying dilation and rotational values. They are mainly implemented by convolving a function with the image to generate Gabor Space [2].For a set of degree of rotations and dilations a set of Gabor kernels are generated. These kernels will extract the 'jets' from the image. [18]

$$J_j(\vec{x}) = \int I(\vec{x}')\psi_j(\vec{x} - \vec{x}')d^2\vec{x}' \tag{1}$$

Convolution with Gabor Kernels to generate wavelet transformed image

$$\psi_j(\vec{x}) = \frac{k_j^2}{\sigma^2}\exp\left(-\frac{k_j^2 x^2}{2\sigma^2}\right)\left[exp(i\overrightarrow{k_j}\vec{x}) - \exp\left(-\frac{\sigma^2}{2}\right)\right] \tag{2}$$

Family of Gabor Kernels for *j* varying from 0 to 39

Here k_j is the wave vector that is restricted by the Gaussian envelope function. For our calculations, 5 different set of frequencies for index $\gamma = 0, 1 \ldots 4$ and 8 sets of orientation directions $\mu = 0, 1, 2 \ldots 7$ are taken [18].

$$\overrightarrow{k_j} = \begin{pmatrix} k_{jx} \\ k_{jy} \end{pmatrix} = \begin{pmatrix} k_\nu \cos\varphi_\mu \\ k_\nu \sin\varphi_\mu \end{pmatrix}, k_\nu = 2^{-\frac{\nu+2}{2}}\pi, \varphi_\mu = \mu\frac{\pi}{8}, \tag{3}$$

Wave vector *kj* for *j* varying from 1 to 39

The width $\frac{\sigma}{k}$ is Gaussian controlled with $\sigma = 2\boldsymbol{\pi}$. The preference of Gabor wavelet transform over normal edge detection and analysis is evident in this case as Gabor filters are much more robust to the data format of biological relevance which in this case is facial segments. Also the robustness is defined as the result of the transform is not susceptible to variation brightness when the Gabor wavelets are considered DC-free [18].

If these jets are normalized then the Gabor wavelet transform is set immune to contrast variations. Still rotation and translation to a small degree does not affect the magnitude of jets but result in high phase variations.

The phase can be used to calculate the degree of displacement between two images and compensate this distortion. [13][12]

Jet Similarity

These phase values have its set of importance in feature matching [9] as:

1. Similar magnitude patterns can be discriminated
2. The phase variation is a measure for accurate jet localization of a feature point.

In order to stabilize the phase sensitive similarity function, the compensation factor needs to be subtracted that nullifies the phase variation in nearby pixel points. For the compensation factor it is assumed that the jets compared in the similarity function

belongs to nearby point hence have a small displacement between them [18]. Thus a small relative displacement d is implemented to generate following phase sensitive similarity function,

$$S_\phi(J,J') = \frac{\sum_j a_j a'_j \cos\left(\phi_j - \phi'_j - \vec{dk_j}\right)}{\sqrt{\sum_j a_j^2 \sum_j a_j'^2}} \tag{4}$$

Similarity Function S for jets including phase

For displacement factor d, the Taylor series expansion is further solved to get reduced to:

$$\vec{d}(J,J') = \begin{pmatrix} d_x \\ d_y \end{pmatrix} = \frac{1}{\Gamma_{xx}\Gamma_{yy} - \Gamma_{xy}\Gamma_{yx}} X \begin{pmatrix} \Gamma_{yy} & -\Gamma_{yx} \\ -\Gamma_{xy} & \Gamma_{xx} \end{pmatrix} \begin{pmatrix} \phi_x \\ \phi_y \end{pmatrix} \tag{5}$$

$$\phi_x = \sum_j a_j a'_j k_{jx} (\phi_j - \phi'_j)$$

$$\Gamma_{xy} = \sum_j a_j a'_j k_{jx} k_{jy}$$

Displacement vector

This equation gives the displacement factor between the jets considering they belong to two neighboring pixel points in the locality [18]. Thus the range of the displacement that can be calculated with it extends to half the wavelength of highest frequency kernel that can be up to 8 pixels for high frequencies. [4][16]

Face Graph Representation

For face graph generation from facial images, a set of 'fiducial points' are decided upon where each fiducial point represents a unique facial feature that will help in generating a representative face graph for that person.

For our analysis the fiducial point chosen were - Iris of left eye, Iris of right eye, nose tip, upper lip tip and chin tip.

Hence a labeled graph will have N = 5 nodes and E = 10 edges connecting between those points. Here an edge e<E connects two nodes n and n'.

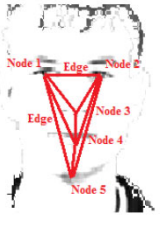


Fig. 3. The face graph that can be generated with the considered set of fiducial points

Face Bunch Graph

For large databases generating a separate face graph for each feature will create an excess database, which can be reduces by bunching data into a facial graph. Hence, a 'face bunch graph' is generated from the individual set that will have a stack of jets at each node that represents a fiducial point. [1]

Face bunch graph is generated for each individual person that represents uniquely the facial characteristics of that person. It has a set of jets from extracted from each model image representative of a particular person.

In our analysis, a small database is taken such that 3-4 face graphs of a person can successfully create a model bunch graph.

5.3 Training

After a bunch graph is generated representative of each individual, each input face segment is used to generate a face graph which is then iteratively matched with each and every of the bunch graph. [18]

Initially a training set of facial images is taken; each image is marked to a corresponding person. This set of image is used to manually generate face bunch graph [8]

For face recognition, fiducial point interior to the face region is important for identification procedure. For this, we choose out 5 fiducial points in the interior region of the face segment.

Graph Similarity Matching

Thus for an image graph G with nodes n = 1, 2…N and edges e=1,2,…E matching is done between the corresponding parameters of the face bunch graph B as[18]:

$$S_{\mathrm{B}}(G^I, B) = \frac{1}{N}\sum_{n} max_m \left(S_\phi(J_n^I, J_n^{Bm})\right) - \frac{\lambda}{E}\sum_{e}\frac{(\Delta\vec{x}_e^I - \Delta\vec{\mathrm{x}}_{\mathrm{e}}^{\mathrm{B}})^2}{(\Delta\vec{\mathrm{x}}_{\mathrm{e}}^{\mathrm{B}})^2} \tag{6}$$

Graph Similarity Measure between image graph and bunch graph

λ decides the relative importance between jets and metric structure at 1 for our analysis. The locality about a fiducial point is taken to be of a range (+/- 5, +/-5) pixels.

5.4 Image Recognition

As in case of our analysis, there is face bunch graph representative of 5 test people. Each of the face bunch graphs is trained with a set of 2-3 image graphs [9].

Our recognizer matches the input image with the entire available face bunch graph and generates a unique similarity measure (Sb1, Sb2…Sb5).

Now this similarity measure value is input to the recognition thresholding limiter that generates binary output '1' or '0'. These outputs (O1, O2…O5) are multiplied by weights (W1, W2…W5) and then summed to generate a recognition index,

$$\mathrm{R=O1*W1+O2*W2+O3*W3+O4*W4+O5*W5,} \tag{7}$$

Recognition Index for example database

If the Recognition index is:

1. '0', then the picture is not a match to the database
2. Same as any single weight(W1, W2,…W5), then it is perfect recognition
3. Sum of two or more weights, then it is over recognition requiring manual correction

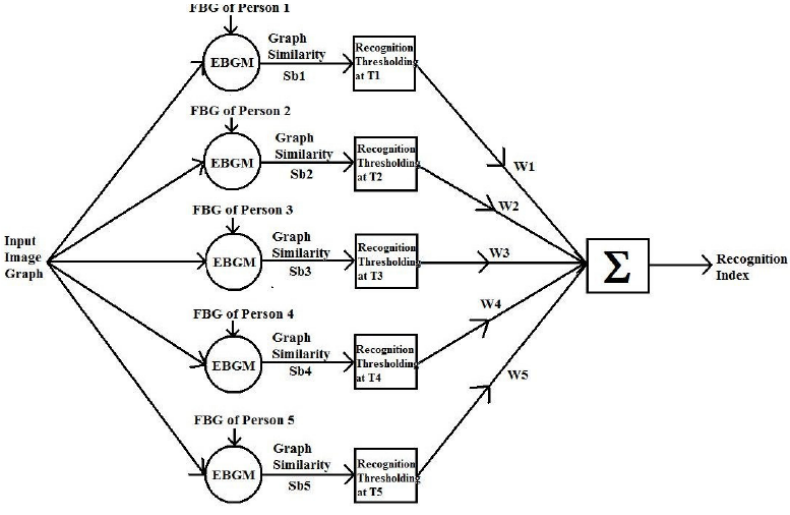


Fig. 4. Recognition network

6 Experimental Result

For result analysis, Face Recognition Database of University of Essex, UK is used which has color images of 395 individuals (male and female) with 20 images per individual of people of various races of age between 18 to 20 years.

To prove the utilization of the algorithm in real life applications where the availability of 20 images per person is improbable, each of the 15 individual Bunch Graph Models were trained using 1 image per individual with manually marked fiducial points.

For initialization of the algorithm, Similarity measure of each bunch graph is tested with 1 same individual image and 1 different individual image. For all the test images (arbitrarily selected) the similarity criteria for the similar images were found to be >0.9985 and that of the dissimilar images were found to be <0.9983 providing a threshold bandwidth of 0.0002.

The testing was performed on a truncated database of arbitrary 15 images per individual for all the 15 individuals. Therefore the total truncated database consisted of 225 facial images.

For best recognition rates the frequency of the Gabor Filters were restricted to $\gamma = \frac{\pi}{4}$ and the orientations were varied from 0, $\frac{\pi}{8}$,…,π.

Table 1. Comparison of Matching Accuracy

Matching Accuracy	Wiskott-EBGM	Skin-Segmentation followed by Wiskott EBGM
University of Essex Database	85%	88%

7 Limitations

The Wiskott-EBGM is prone to translational and rotational distortions, and though the rotational distortions were considered zero degrees, the translation distortion was minimized to 5 pixels to reduce error margin, which reduces the robustness of the algorithm.

The Matching Rate was reduced due to over identification, which can be controlled for large databases using continuous user feedback.

8 Conclusion and Future Work

In this paper, a two-step facial recognition algorithm is implemented for analysis. Instead of directly implementing EBGM on images, EBGM algorithm is tested on skin segmented images. As the skin segmented images only have the necessary facial data, without background noises, it reduces the over identified Gabor features of the background and increases the efficiency.

The enhancement of the Matching Accuracy is only by 3% because though the background noise could be removed, the intra-facial region noises could not be eliminated. Also, as per our objective the modified two-step EBGM algorithm can successfully identify multiple facial regions in input images with multiple faces and extract them separately for matching.

For further enhancement of Matching Accuracy, the traditional Wiskott EBGM algorithm can be optimized to extract relevant facial features.

References

1. Beymer, D.: Face recognition under varying pose. In: Proc. IEEE Computer Vision and Pattern, pp. 756–761 (1994)
2. Daugman, J.G.: Complete discrete 2-D Gabor transform by neural networks for image analysis. IEEE Trans. on Acoustics, Speech and Signal Processing (1988)
3. DeValois, R.L.: Spatial Vision. Oxford Press (1988)
4. Fleet, D.J.: Computation of component image velocity from local phase information. Int'l J. of Computer Vision, 77–104 (1990)
5. Goldstein, A., Harmon, L., Lesk, A.: Identification of Human Faces. IEEE Proceedings 59(5), 748–760 (1971)
6. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice Hall, New Jersey

7. Jean-Marc Fellous, N.K.: Face Recognition by Elastic Bunch Graph Matching. In: Jain, L. (ed.) Intelligent Biometric Techniques in Fingerprint and Face Recognition, pp. 355–396. CRC Press (1999)
8. Kruger, N.P.: Determination of face position. Image and Vision Computing (1997)
9. Lades, M.V.: Distortion invariant object recognition in the dynamic link architecture. IEEE Trans. on Computers, pp. 300–311 (1993)
10. Mohsin, W., Ahmed, N., Mar, C.-T.: Face Detection Project (2003)
11. Movellan, J.: Introduction to Gabor Filters. NSTC Subcommittee on Biometrics (n.d.) Retrieved from, http://www.biometrics.gov/Documents/facerec.pdf
12. Pollen, D.A.: Phase relationship between adjacent simple cells in the visual cortex. Science, 1409–1411 (1981)
13. Potzsch, M.K.: Improving object recognition by transforming Gabor Filter responses. Network: Computation in Neural Systems, 341–347 (1997)
14. Saveliev, P. (n.d.): Euler Number - Computer Vision Primer. Retrieved from Intelligent Perception, http://inperc.com/wiki/index.php?title=Euler_number
15. Sirovich, L., Kirby, M.: A Low-Dimensional procedure for Characterization of Human Faces. Optical Soc. Am. A 4(3), 519–524 (1987)
16. Theimer, W.M.: Phase-based binocular vergence control and depth reconstruction using active vision. In: CVGIP: Image Understanding, pp. 343–358 (1994)
17. Turk, M., Pentland, A.: Face Recognition Using Eigenfaces. IEEE Proceedings, 586–591 (1991)
18. Wiskott, L.F.-M.: Face recognition by elastic bunch graph matching. IEEE Trans. on Pattern Analysis and Machine Intelligence, 775–779 (1997)
19. Xiong, Z.: An Introduction to Face Detection and Recognition

A Study of Prosodic Features of Emotional Speech

X. Arputha Rathina[1], K.M. Mehata[1], and M. Ponnavaikko[2]

[1] Department of Computer Science and Engineering, B.S. Abdur Rahman University, Vandalur, Chennai, India
[2] SRM University, Kattankulathur, Chennai, India

Abstract. Speech is a rich source of information which gives not only about what a speaker says, but also about what the speaker's attitude is toward the listener and toward the topic under discussion—as well as the speaker's own current state of mind. Recently increasing attention has been directed to the study of the emotional content of speech signals, and hence, many systems have been proposed to identify the emotional content of a spoken utterance.

The focus of this research work is to enhance man machine interface by focusing on user's speech emotion. This paper gives the results of the basic analysis on prosodic features and also compares the prosodic features of, various types and degrees of emotional expressions in Tamil speech based on the auditory impressions between the two genders of speakers as well as listeners. The speech samples consist of "neutral" speech as well as speech with three types of emotions ("anger", "joy", and "sadness") of three degrees ("light", "medium", and "strong"). A listening test is also being conducted using 300 speech samples uttered by students at the ages of 19 -22. The features of prosodic parameters based on the emotional speech classified according to the auditory impressions of the subjects are analyzed. Analysis results suggest that prosodic features that identify their emotions and degrees are not only speakers' gender dependent, but also listeners' gender dependent.

1 Introduction

One can take advantage of the fact that changes in the autonomic nervous system indirectly alter speech, and use this information to produce systems capable of recognizing affect based on extracted features of speech. For example, speech produced in a state of fear, anger or joy becomes faster, louder, precisely enunciated with a higher and wider pitch range. Other emotions such as tiredness, boredom or sadness, lead to slower, lower-pitched and slurred speech. Emotional speech processing recognizes the user's emotional state by analyzing speech patterns.

Previous researches have engaged in the study of emotions, in how to recognize them automatically from speech and they have tried to incorporate this technology into real world applications. However, it has been difficult to find the features that describe the emotional content in speech. It has not been reached a reliable set of features for discriminating emotional states in spontaneous speech [1]. We can find information associated with emotions from a combination of prosodic and spectral information. Much of the work done to date has been focused on features related to prosodic aspects.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 41–49.
springerlink.com

As information communication technology (ICT) advances, there are increasing needs for better human-machine communication tools. Expressive speech is more desirable than non-expressive speech as a means of man-machine dialog. However, the capability of synthesizing expressive speech including emotional speech is currently not high enough to match the needs. We have to explore features from natural speech to achieve a method for a variety of expressive-speech synthesis. Among expressive speech, we have so far placed a focus on emotional speech.

New applications such as speech-to-speech translation, dialogue or multimodal systems demand for attitude and emotion modeling. Humans would choose different ways to pronounce the same sentence depending on their intention, emotional state, etc. Here we present a first attempt to identify such events in speech. For that purpose we will try to classify emotions by means of prosodic features.

Prosody is a rich source of information in speech processing. Prosody has long been studied as an important knowledge source for speech understanding and also considered as the most significant factor of emotional expressions in speech [1, 3]. In recent years there has been a large amount of computational work aimed at prosodic modeling for automatic speech recognition and understanding. This paper places a focus on the study of the basic correlation between Prosodic features like Pitch contour, energy contour and utterance timing and emotional characteristics.

The emotion types like “anger”, ”sad”, ”happy” and “neutral” are used for the analysis. Speakers are the students in the age group of 19 -22. A minute approach instead of generally investigating various types of emotional expressions is taken into consideration [4]. The degree of emotion are categorized into “Neutral”, “Low” and “Strong”,[5] and the prosodic features of each category have been analyzed.

In this analysis so far, the type and degree of each emotion has been determined by the speakers themselves. In conversational communication, however, a speaker’s emotion inside his/her mind is not necessarily reflected in his/her utterances, nor is exactly conveyed to the listener as the speaker intended. The purpose of our study therefore is to clarify quantitatively (1) how much the speaker’s internal emotion (speaker’s intention) is correctly conveyed to the listener and further, (2) what type of expression is able to convey the speaker’s intention to the listener correctly. As the style of emotional expressions is gender-dependent, the gender features are also taken into consideration.

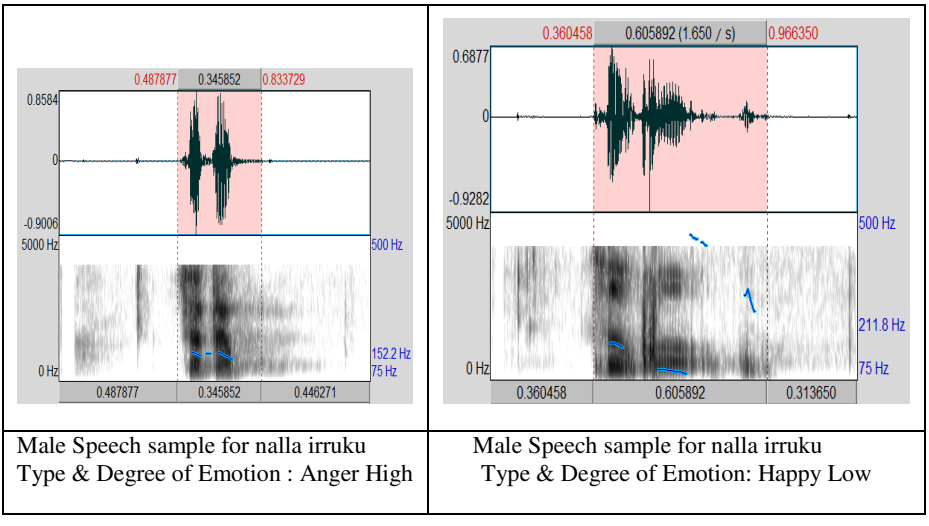
We first conducted a listening test to examine how much the speaker‘s intended emotions agreed with the listeners auditory impressions, using 130 word speech sample uttered by the college students at the age of 19-22 year old. The test results show that the subjects need not necessarily perceive emotional speech as the speakers intended to express.

From these results, we therefore analyzed the features of prosodic parameters based on the emotional speech classified according to the auditory impressions of the subjects. Prior to analysis, we calculated an identification rate of each type and degree of emotion, which was rate of the number of identifying as the specific type and degree of emotion to the total number of listeners. We selected 5 speech samples whose identification rates ranked the top 4 for each type and degree of emotion.

2 Implementation

2.1 Speech Samples

The speakers are college students in the age groups of 19 to 22. As speech samples, we use 2- Tamil words: “nalla iruku”, and ”ennada panura”. The types of emotions are “anger”, “happy”, and “sad”. Each word is uttered with following 2 degrees of emotions: “high”, and “low” and the speech samples are given below in figure 1.



Male Speech sample for nalla irruku Type & Degree of Emotion : Anger High	Male Speech sample for nalla irruku Type & Degree of Emotion: Happy Low

Fig. 1. Speech Samples

2.2 Prosodic Feature Parameters

Prosodic-feature parameters used in this work are Pitch contour, Utterance Timing and Energy contour. We did not use the speech power because the distances between the subjects and the microphone varied largely by their body movements during recording and we could not collect reliable power data.

These features have been extracted by means of MATLAB programs using Signal processing toolbox.

- **Pitch Contour**

The *pitch signal*, also known as the glottal waveform, has information about emotion, because it depends on the tension of the vocal folds and the subglottal air pressure. The pitch signal is produced from the vibration of the vocal folds. Two features related to the pitch signal are widely used, namely the pitch frequency and the glottal air velocity at the vocal fold opening time instant. For pitch, female pitch voice ranges from 150-300, for male pitch voice ranges from 50-200 and for child pitch voice ranges from 200-400. The time elapsed between two successive vocal fold openings is

called *pitch period T*, while the vibration rate of the vocal folds is the *fundamental frequency of the phonation F*0 or *pitch frequency*. The glottal volume velocity denotes the air velocity through glottis during the vocal fold vibration. High velocity indicates music like speech like joy or surprise. Low velocity is in harsher styles such as anger or disgust. The pitch estimation algorithm used in this work is based on the autocorrelation and it is the most frequent one. A widely spread method for extracting pitch is based on the *autocorrelation of center-clipped* frames. The signal is low filtered at 900 Hz and then it is segmented to short-time frames of speech *fs*(*n*;*m*). The clipping, which is a nonlinear procedure that prevents the 1st formant interfering with the pitch, is applied to each frame *fs*(*n*;*m*) yielding

$$\hat{f}_s(n;m) = \begin{cases} f_s(n;m) - C_{thr} & \text{if } |f_s(n;m)| > C_{thr} \\ 0 & \text{if } |f_s(n;m)| < C_{thr} \end{cases},$$

where *Cthr* is set at the 30% of the maximum value of *fs*(*n*;*m*). After calculating the short-term autocorrelation

$$r_s(\eta;m) = \frac{1}{N_w} \sum_{n=m-N_w+1}^{m} \hat{f}_s(n;m)\hat{f}_s(n-\eta;m),$$

where η is the lag, the pitch frequency of the frame ending at *m* can be estimated by

$$\hat{F}_0(m) = \frac{F_s}{N_w} \operatorname{argmax}_{\eta}\{|r(\eta;m)|\}_{\eta=N_w\,(F_l/F_s)}^{\eta=N_w\,(F_h/F_s)},$$

where *Fs* is the sampling frequency, and *Fl*, *Fh* are the lowest and highest perceived pitch frequencies by humans, respectively. The maximum value of the autocorrelation (max*{|r*(*η*;*m*)*|}η=Nw* (*Fh/Fs*) *η=Nw* (*Fl/Fs*)) is used as a measurement of the glottal velocity during the vocal fold opening.

- **Utterance Timing**

There are two ways of extracting the utterance timing features. The first one is based on the length of syllables and the second one is based on the duration of pauses into the utterance voice periods. [10] To calculate the former type, we have to use the technique where syllables are detected automatically without needing a transcription. After detecting syllables, the total sounding time for every recording has to be calculated. The speech rate for every recording is obtained dividing the total amount of detected syllables by the total sounding time of the recording. From this procedure we can calculate: speech rate, syllable duration mean and syllable duration standard deviation.

The latter type of Utterance Timing features is extracted from the calculation of the silences and voice period's durations in the utterance. This work uses this method and the features extracted are: pause to speech ratio, pause duration mean, pause duration standard deviation, voice duration mean and voice duration standard deviation. *Figure 2 gives the details of the Utterance timing features extracted for a sample speech.*

- **Energy Contour**

Energy is the acoustic correlated of loudness; their relation is not linear, because it strongly depends on the sensitivity of the human auditory system to different frequencies. Coupling of the loudness perception with the acoustic measurement is as complex as the coupling of the tone pitch perception and the computable F0. The sensation of loudness is both dependent on the frequency of the sound and on the duration, and the other way round, pitch perception depends on the loudness [10]. The short time speech energy can be exploited for emotion detection, because it is related to the arousal level of emotion. The short time energy of the speech frame ending at m is

$$E_s(m) = \frac{1}{N_w} \sum_{n=m-N_w+1}^{m} |f_s(n;m)|^2.$$

Figure 2 shows the combination of all the above three prosodic feature extraction of a given speech sample.

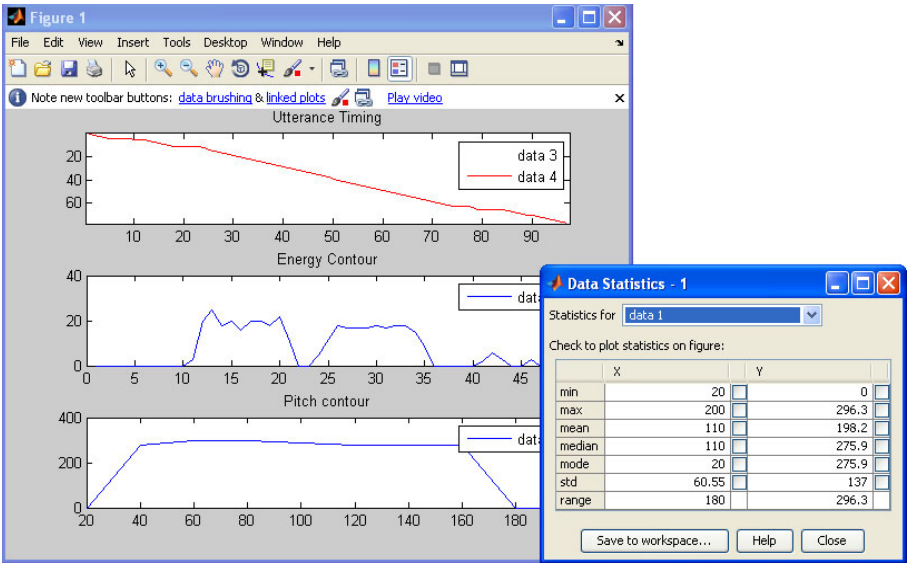


Fig. 2. Extraction of Pitch, Energy Contour and Utterance Timing from a sample speech

3 Experimental Results

3.1 Experimental Conditions

Listening tests have been conducted to investigate whether or not there are Listeners and speakers gender dependent difference in the prosodic features of speech samples that have the same auditory impression on the type and degree.

In the listening tests, speech samples were presented to the subjects in random order. There were 6 dummy samples ahead and 100 test samples. We conducted two sessions for the purpose of cancelling the order effect. In the second session, the speech samples were presented to the subjects in the reverse order of those presented in the first session. Fifty subjects used a headphone of the same maker and the same sound pressure. Among them, 25 subjects were male at the ages of 20 and 21 years old and 25 subjects were female college students at the ages of 19 and 20 years old, both with a normal auditory capacity.

3.2 Identification Rate

To quantify the strength of listener's auditory impressions, an "identification rate r" [11] was introduced. We defined "an identification number" as the number of listeners' identification regardless of the type and degree of emotion that speaker intended to express. In the same way, we defined "*an identification rate r*" as a rate of the identification number to the total number of listeners.

Table 1 lists the top 5 speech samples in identification rate for each gender combination of speakers and listeners extracted from all types and degrees of emotional speech.

Table 1. The rank of Identification rate r for all speech

Speaker	Rank	Male listener	Female Listener
		Degree & Emotion (Id. Rate r %)	
Male	1	Anger high(100.0)	Sad high(100.0)
	2	Happy high(94.0)	Sad low(98.0)
	3	Happy low(92.0)	Happy high(97.0)
	4	Anger low(91.0)	Happy low(94.0)
	5	Sad high (86.0)	Anger high(94.0)
Female	1	Anger low(89.0)	Anger low(99.0)
	2	Happy low(87.0)	Happy low(90.0)
	3	Sad low(83.0)	Happy high(82.0)
	4	Anger high(80.0)	Sad high(80.0)
	5	Sad high(79.0)	Sad low(78.0)

In the case of speech uttered by Female speakers, the emotions and degrees that had the top 5 identification rates are "Anger low" and "Happy low" perceived by both male and female listeners. In the case of speech uttered by male speakers, on the other hand, speech sample perceived as "Anger low" and "Happy low" are not included in the emotions and degree that had top 5 identification rates. The emotion and degree that had the top 5 identification rates are "anger high" for male listeners and "sad high" for female listeners.

3.3 Database Samples and Results

The Database contains 300 speech samples, 150 each for the two Tamil words: "nalla iruku", and"ennada panura". Each word is uttered in different types of emotions: "anger", "happy", "sad" and in different degrees: "High" and "Low". Using MATLAB programs we have extracted the prosodic features for all the samples. Thus the Database contains all the samples and their corresponding prosodic features. We computed and compared the average prosodic features from all the 300 samples and the results are given below:

Figure 3 gives the comparison of the average Pitch range all the Male and Female speech samples present in the database. Figure 4 gives the comparative chart of different emotions of both male and female speech samples under different degrees. Figure 5 says that Female pitch is greater than Male but the energy counter and utterance timings of male are slightly greater than female.

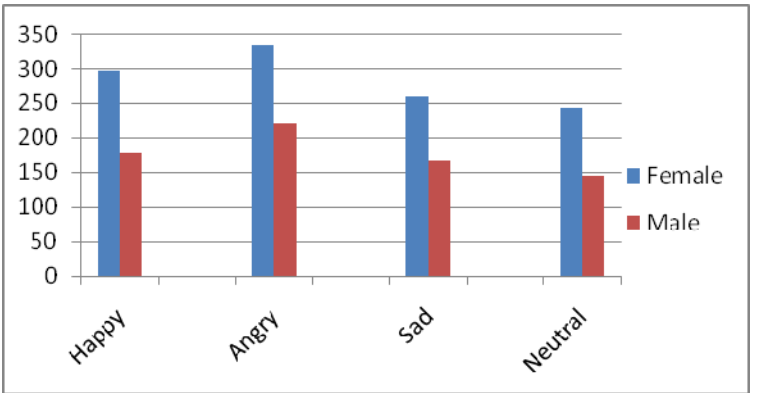


Fig. 3. Average Pitch Comparison of Male and Female speakers'

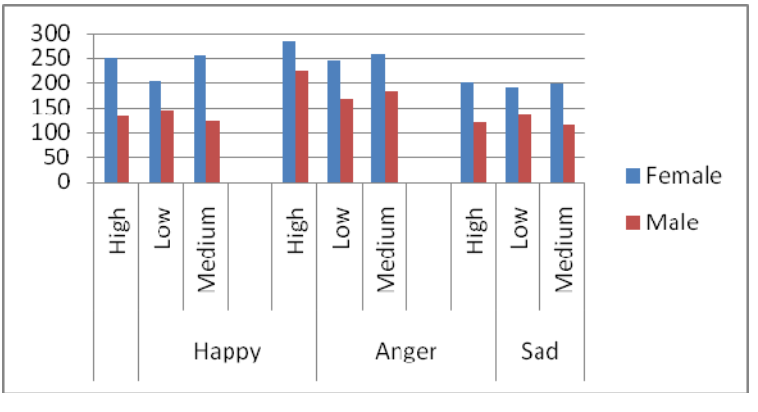


Fig. 4. Comparisons' of all emotional in different degrees

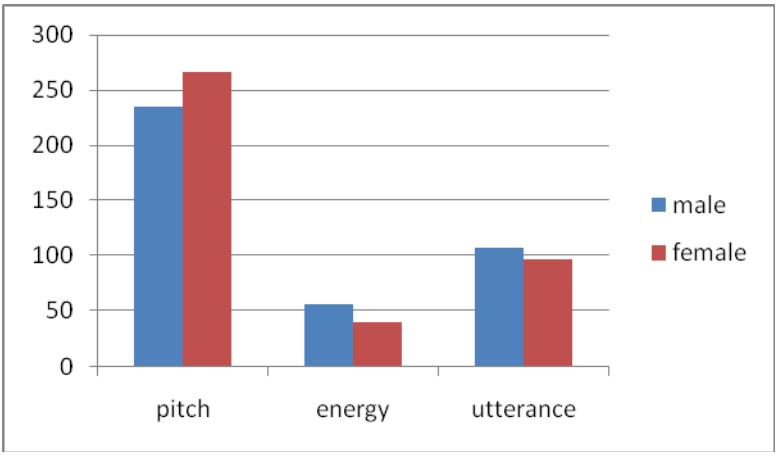


Fig. 5. Comparison of Pitch Contour, Energy and Utterance Timing

4 Conclusion

The test results have suggested that there are Listeners as well as Speakers gender dependent differences in prosodic features to identify the type and degree of emotions.

Analysis results are summarized as follows: 1. For anger speech uttered by female speakers increases compared to that for neutral speech and significant difference has been observed from male speech. 2. Prosodic features that characterize their emotions' are speaker gender- dependent. 3. Pitch for all emotion of female speech increases, and also significant difference has been observed from male speech.

Experimental results show the possibility in which our approach is effective for enhancement in Human Computer Interactions. Future works of our research are the following. We have to collect synthetic speech and put emotion labels on them. We have to reconsider how to estimate emotion in speech based on the results of experiments. For example, which features do we focus on? Which combination features and learning algorithms do we use? We have to reconsider evaluation of our approach and do it, too. People express and estimate more than one emotion in human speech. So we should think processing multi emotions in speech to develop better human computer interaction.

References

[1] Kuremastsu, M., et al: An extraction of emotion in human speech using speech synthesize and classifiers foe each emotion. International Journal of Circuits Systems and Signal Processing (2008)

[2] Adell, J., Bonafonte, A., et al.: Analysis of prosodic features: towards modeling of emotional and pragmatic attributes of speech. In: Proc. Natural Lang. Proc. (2005)

[3] Hashizawa, Y., Takeda, S., Hamzah, M.D., Ohyama, G.: On the Differences in Prosodic Features of Emotional Expressions in Japanese Speech according to the Degree of the Emotion. In: Proc. 2nd Int. Conf. Speech Prosody, Nara, Japan, pp. 655–658 (2004)
[4] Takeda, S., Ohyama, G., Tochitani, A.: "Diversity of Prosody and its Quantative Description" and example: analysis of "anger" expression in Japanese Speech. In: Proc. ICSP 2001, Taejon, Korea, pp. 423–428 (2001)
[5] Hamzah, M.D., Muraoka, T., Ohashi, T.: Analysis of Prosodic features of Emotional Expressions in Noh Farce speech according to the Degree of Emotions. In: Proc. 2nd Int. Conf. Speech Prosody, Nara, Japan, pp. 651–654 (2004)
[6] Espinosa, H.P., Reyes Garcıa, C.A., Pineda, L.V.: Features Selection for Primitives Estimation on Emotional Speech. In: ICASSP (2010)
[7] Zhang, S., Lei, B., Chen, A., Chen, C., Chen, Y.: Spoken Emotion Recognition Using Local Fisher Discriminant Analysis. In: ICSP (2010)
[8] Liscombe, J.: Detecting Emotions in Speech: Experiments in three domains. In: Proc. of the Human Lang. Tech Conf. of the North America Chapter of ACL, pp. 234–251 (June 2006)
[9] Dumouchel, P., Boufaden, N.: Leveraging emotion detection using emotions form yes-no answers. In: Interspeech (2008)
[10] Galanis, D., Darsinos, V., Kokkinakis, G.: Investigating Emotional Speech Parameters for Speech Synthesis. In: ICECS (1996)
[11] Adell, J., Bonafonte, A., et al.: Analysis of prosodic features: towards modeling of emotional and pragmatic attributes of speech. Proc. Natural Lang. Proc. (2005)

Gender Classification Techniques: A Review

Preeti Rai and Pritee Khanna

PDPM Indian Institute of Information Technology,
Design & Manufacturing Jabalpur, India

Abstract. Face is one of the most important biometric traits. By analyzing the face we get a lot of information such as age, gender, ethnicity, identity, expression, etc. A gender classification system uses face of a person from a given image to tell the gender (male/female) of the given person. A successful gender classification approach can boost the performance of many other applications including face recognition and smart human-computer interface. This paper illustrates the general processing steps for gender classification based on frontal face images. In this study, several techniques used in various steps of gender classification, i.e. feature extraction and classification, are also presented and compared.

Keywords: Biometrics, feature extraction, classifier.

1 Introduction

Visualinformation plays an important role when people communicate with each other. When we look at a person's face, not only we discern who he/she is, but also process other information about him/her, such as gender, ethnicity, age as well as the current state of mind through expressions. Gender classification is to tell the gender of a person according to his/her face. It is an easy job for humans, but a challenging one for computers. Gender classification could be of important value in human–computer interaction, such as personal identification. Also, it is a useful preprocessing step for face recognition. A computer system with the capability of gender classification has a wide range of applications in basic and applied research areas, including man-machine communication, security, law enforcement, demographics studies, psychiatry, education and telecommunication, etc. The field of face recognition has been explored by many researchers. But in gender recognition or classification only a few works have been reported. The face image is used for classifying the gender, so the gender classification process can make face recognition twice as fast by reducing the search time for recognizing the person.

Earlier, the gender classification techniques were based on neural network. A two-layer SEXNETisdeveloped with 30 by 30 pixel face (Gollomb et al. 1991) samples. The Support Vector Machine is used to classify gender with low-resolution 21*12 “thumbnail” faces (Moghaddam et al. 2002). An Automatic Real-Time Gender Classification system isintroduced thatwasbased on LUT-Adaboost method (Wu et al. 2003). The objective of this paper is to summarize and compare various gender classification techniques and the related future research issues.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 51–59.
springerlink.com

The rest of this paper is organized as follows: Section 2 discusses the overview of gender classification systems; Section 3 gives details of the feature extraction methods; inSection 4, the gender classifiers is described; the comparison of various gender classification techniques is presented in section 5; finally section 6 concludes the work and its future scope.

2 System Overview

Fig. 1 shows the gender classification system which consists of feature extraction and classifier module. In the training phase, feature extraction module reduces the data by measuring certain "features" or "properties" of the training face images that are useful for classification. After this features are stored in the database. While in the testing phase, features of the test face image are extracted and these extracted features are used by the classifier to classify the image with the help of the database which is created during the training phase and makes the final decision.

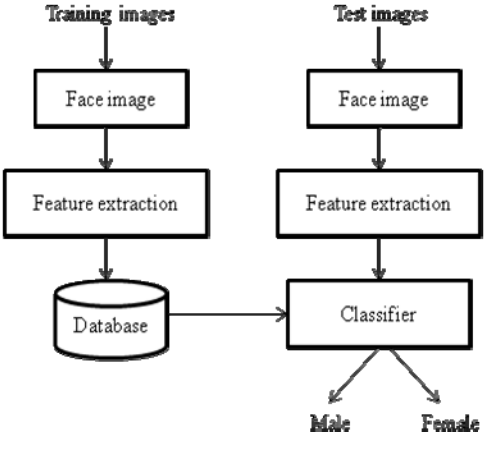


Fig. 1. Gender classification system

3 Feature Extraction

Features should be easily computed, robust, insensitive to various distortions and variations in the images, and rotationally invariant. Based on the type of features used, previous studies can be broadly classified into two categories:

- Appearance feature-based (global)
- Geometrical feature-based (local)

The first approach finds the decision boundary directly from training images, while other approach is based on geometric features such as eyebrows thickness, nose

width, etc. The first strategy usually considers an image as a high-dimensional vector and extracts features from its statistical information, without relying on knowledge about the object of interest. Typically, this approach is holistic and has the advantage of being fast and simple. However, such a representation can become unreliable when local appearance variations occur. In the second strategy, some apriori knowledge was applied, and facial geometry features such as the eyes, nose, and mouth are extracted. In facial feature detection methods, appearance based approaches are considered more often than the geometrical based methods.

In the system proposed by Wiskott et al., face images are represented as graphs labeled with topographical information and local templates (Wiskott et al. 1995). Gender classification is done by comparing the graphs using a similarity function which makes the system efficient in recognizing the face. (Lyons et al. 2000) used Gabor filter to extract features from the face and applied Principle Component Analysis (PCA) for dimension reduction. In addition to the PCA, the Independent Component Analysis (ICA) (Graf et al. 2002), local linear embedding (Buchala et al. 2004) and curvilinear component analysis (Jain and Huang 2004) have also been applied for extracting the features from the face.

In another approach, the features are extracted from the face image using PCA and genetic algorithm, for gender classification (Sun et al. 2002).

A novel method proposed for classifying the gender used Local Binary Pattern (LBP) for face feature extraction (Sun et al. 2006). (Lian and Lu 2006), also experimented with LBPs and achieved good results. (Baluja and Rowley 2007) presented a method, based on Adaboost for identifying the gender from a low resolution image.

Different feature extraction methods (Gabor Wavelets, Haar-Like Wavelets, Principal Component Analysis (PCA), and Independent Component Analysis (ICA))werecompared in(Lu and Lin 2007). Their experimental results on FERET database of frontal facial images showed that the Gabor features of face images method hadachieved the best performance. As the Gabor filters can extract the face features with different orientations and scales, it has strong representation ability. The mean Adaboost and local binary pattern methods are used for extracting the facial features (Makinen and Raisamo 2008). APixel-Pattern-Based Texture Feature (PPBTF) is proposed for gender recognition (Lu et al. 2008). In this method (PPBTF), Adaboost and SVM are used to classify gender and achieved the classification above 90%.

A fusion based method is also proposed for gender classification. Experiments on FERET images showed that fusion outperforms individual features, and using CAS-PEAL images it was also shown that the fused results improve on the full face approach (Lu and Shi 2009). A 2DPCA method isemployedfor extracting features from the face for gender classification (Bui et al. 2010). A novel texture descriptor Local Directional Pattern (LDP) is used to represent facial image for gender classification. In this descriptor the face area is divided into small regions, for each region LDP histograms are extracted and concatenated into a single vector to efficiently represent the face image(Jabid 2010).

In our previous work (Rai and Khanna 2010) the combination of radon and wavelet transforms are used to extract features of the face with different orientations with respect to the illumination and expression changes.

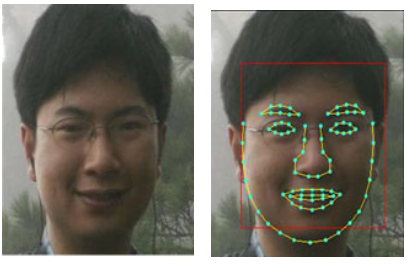
(Alexander 2010) used amultiscale decision fusion approachto recognize the gender. (Yan 2011) applied 2D Gabor transform for gender recognition. (Li et al. 2010) utilized the non-tensor product prewavelets to extract the face features for classifying the face.

A hybrid approachcombining PCA and SFS (Sequential Forward Selection) for face feature extraction is given by (Basmacei 2011). All the approaches discussed above are appearance based methods.

In Geometrical based approach, the Active Appearance Model (AAM) is used to locate 83 landmarks, got 3403 geometry features, from which 10 most significant features were picked, normalized and fused with the appearance features for gender recognition (Xu at el. 2008).

(Brunelli and Poggio 1992)have extracted the following three features among the sixteen geometric features, for their classification system: (a) distance of eyebrow from eyes (b) eyebrows thickness and (c) nose width.

(Samal et al. 2007) experimented on 406 geometry features. (Xia et al. 2009) worked on three dimension facial features. Their methods were based on geometrical feature measurement of three dimensional faces. The geometrical features are obtained from facial image and canny edge operator isappliedto locate the position of eyes, mouth and nose(Ramesh et al 2009, Ramesh et al 2010).An Active Appearance Model (AAM) (Xu at el. 2008, Makinen and Raisamo 2008) is applied on the face to get 83 landmarks from face image as shown in Fig. 2.



(a) (b)

Fig. 2. Landmarks located by AAM [Xu at el. 2008]

4 Gender Classifiers

Gender classification is a typical binary classification problem. Once all the facial features are extracted, a classifier is trained and tested which can classify input vector as a male or female. (Makinen and Raisamo 2008) have given the detailed explanation of the gender classifiers which is mentioned below.

4.1 Multilayer Neural Network

An Artificial Neural Network is an information processing system that is inspired by the way biological nervous system, such as brain, processes information. Neural Network has ability to learn how to do tasks based on the data given for training or initial experience. The number of the input nodes in this network is equal to the amount of the pixels in the resized face image. Then, the resized face images are histogram equalized. The intensity values from the histogram equalized images are scaled in a range of -0.5 to 0.5. The network produces output between -0.5 to 0.5. The negative value defines female and positive value a male class. The Neural Network was trained using the standard back propagationalgorithm (Makinen and Raisamo 2008). Neural Network approach requires to findout the right Neural Network architecture (i.e. number of layers, hidden units, etc.) and parameters (learning rates, etc.).

The Neural Network has shown that even a very low resolution image can be used for gender recognition and achieved 93% classification rate (Tamura et al. 1996).

4.2 Support Vector Machine (SVM)

The SVM is a recently developed learning method for pattern classification and regression. The basic idea of the SVM (Sonka and Hlavac 2007) is to find the optimal hyperplane that has the maximum margin of separation between the classes, while having minimum classification errors for linear separable data as shown in Fig. 3. For linearly non-separable data, kernel function maps the input sample to a higher dimensional feature space where a linear hyperplane can be found. Many kernel functions are there but radial basic function (RBF) and polynomial kernels are probably the most used ones (Makinen and Raisamo 2008). It has been observed that accuracy of SVM is better than all other classifiers for low resolution images.

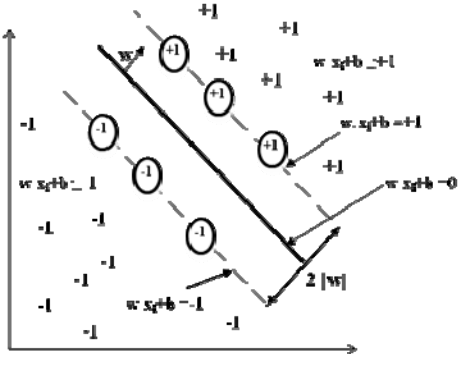


Fig. 3. Hyperplane drawn by the SVM

4.3 DiscreteAdaboost

The specific features are selected using the Adaboost algorithms for gender classification (Freund and Schapire 1996). The weak classifiers that were used with the selected features together form the strong classifier. (Makinen and Raisamo 2008) used haar like features with three kinds of weak classifiers (threshold, mean, LUT (Look-Up Table)). A novel Look-Up Table (LUT) weak classifier based Adaboost approach to learn the gender is presented in (Wu et al. 2003). Threshold weak classifier used in (Freund and Schapire 1996, Shakhnarovich et. al 2002) and LUT both are based on haar like features via integration image, but the latter has stronger ability to model the complex distribution of training samples.

4.4 Other Classifiers

Ak- Nearest Neighbor **(**kNN) is a method for classifying test samples based on nearby training examples in the feature space(Rai and Khanna 2010).Linear Discriminat Analysis (LDA) is also used to classify gender.It finds the direction, along which the classes (Male/Female) are most effectively separated (Jain and Huang 2004, Lyons et al. 2000).

5 Comparison of Various Gender Classification Methods

Classifiers discussed above for gender classification are based either on appearance or feature. The Neural Network and SVM are appearance based while Adaboostisfeature based.

Table 1 gives the classification rates of various combinations of feature extraction methods and classifiers as discussed above.The most popular classifiers are different

Table 1. Classification rate of various gender classification methods

Feature Extraction Techniques	Classifier	Classification rate %
Low resolution image (Tamuraet al. 1996)	Neural Network	93
Gabor+PCA (Lyons et al. 2000)	LDA	92
LBP (Sun et al. 2006)	Adaboost	95.75
PPBTF (Lu et al. 2008)	Adaboost	96.1
PCA + Genetic (Sun et al. 2002)	SVM	95.3
LBP (Lian and Lu 2006)	SVM	96.75
Gabor((Lu et al. 2008)	SVM	85.6
PCA (Lu and Shi 2009)	SVM	92.78
2DPCA (Lu and Shi 2009)	SVM	95.33
Radon and Wavelet (Rai and Khanna 2010)	kNN	84.89

neural network architectures, SVM with different kernel function and boosting methods (primarily Adaboost), which are shown to give better performance for the problem of face gender recognition. Most of the research works show that SVM with RBF (Radial Basic function) should be used to discriminate genders because gender classification is a non linear separable problem.

The previous work analysis also shows that SVM achieves the highest classification rate but has low speed of computation, while in Adaboost and Neural Network one can get high speed of computation but suffers from poor classification rate.

6 Conclusion

This paper attempts to provide a review of research on gender classification. The paper discusses different methods used for feature extraction and gender classification. On the basis of our understanding of various methods, a combination of various techniques gives a better way to utilize available domain knowledge to make automatic and accurate decisions.

Although significant progress has been made in the last decade,but still some work has to be done in this field to design a robust gender classification system, which should be effective under varying lighting condition, orientation, pose, partial occlusion, expressions, etc. In future, robust gender classifier systems based on a face can be designed and optimized so that the time consumed for the feature extraction and classification is minimized with increased system accuracy. The optimization is not a separate step; it can be combined with the steps of the gender classification process, so that classification error rate can be further reduced.

References

1. Golomb, B.A., Lawerence, D.T., Sejnowski, T.J.: Sexnet: A Neural Network Identifies Sex from Human Faces. In: Advances in Neural Information Processing System, pp. 572–577 (1991)
2. Brunelli, R., Poggio, T.: Hyperbf Networks for Gender Classification. In: Proc. DARPA Image Understanding Workshop, pp. 311–314 (1992)
3. Wiskott, L., Fellous, J.M., Krüger, N., Von der Malsburg, C.: Face Recognition and Gender Determination. In: Proc. of Int. Workshop of Automatic Face and Gesture Recognition, pp. 92–97 (1995)
4. Freund, Y., Schapire, R.E.: Experiments with a New Boosting Algorithm. In: Proc. of 13th Int. Conf. on Machine Learning, pp. 148–156 (1996)
5. Tamura, S., Kawai, H., Mitsumoto, H.: Male/Female Identification from 8 *6 Low Resolution Face Images by Neural Networks. Pattern Recognition 29(2), 331–335 (1996)
6. Lyons, M., Budynek, J., Plante, A., Akamatsu: Classifying Facial Attributes Using a 2D Gabor Wavelet Representation and Discriminate Analysis. In: Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 202–207 (2000)
7. Shakhnarovich, G., Viola, P., Moghaddam, B.: A Unified Learning Framework for Real Time Detection and Classification. In: IEEE Conf. on AFG (2002)

8. Moghaddam, B., Yang, M.H.: Gender Classification with Support Vector Machines. IEEE Trans. on PAMI 24(5), 707–711 (2002)
9. Sun, Z., Bebis, G., Yuan, X., Louis, S.J.: Genetic Feature Subset Selection for Gender Classification: A Comparison Study. In: Proc. of IEEE Workshop on Applications of Computer Vision, pp. 165–170 (2002)
10. Graf, A., Wichmann, F.: Gender Classification of Human Faces. In: Proc. of the Int. Workshop on Biologically Motivated Computer Vision, pp. 491–500 (2002)
11. Wu, B., Ai, H., Huang, C.: Real Time Gender Classification. In: 3rd Int. Symposium on Multi-spectral Image Processing and Pattern Recognition, pp. 498–503 (2003)
12. Buchala, S., Davey, N., Frank, R., Gale, T.: Dimensionality Reduction of Face Images for Gender Classification. Technical Report 408, Department of Computer Science, the University of Hertfordshire, UK (2004)
13. Jain, A., Huang, J.: Integrating Independent Components and Linear Discriminant Analysis for Gender Classification. In: Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp. 159–163 (2004)
14. Sun, N., Zheng, W., Sun, C., Zou, C., Zhao, L.: Gender Classification Based on Boosting Local Binary Pattern. In: Wang, J., Yi, Z., Żurada, J.M., Lu, B.-L., Yin, H. (eds.) ISNN 2006. LNCS, vol. 3972, pp. 194–201. Springer, Heidelberg (2006)
15. Lian, H.C., Lu, B.L.: Multi-View Gender Classification Using Local Binary Patterns and Support Vector Machines. In: Proc. 3rd Int. Symposia. on Neural Networks, Chengdu, China, pp. 202–209 (2006)
16. Baluja, S., Rowley, A.H.: Boosting Sex Identification Performance. Int. J. of Computer Vision 71(1), 111–119 (2007)
17. Samal, A., Subramani, V., Marx, D.: Analysis of Sexual Dimorphism in Human Faces. Visual Communication and Image Representation 18(6), 453–463 (2007)
18. Lu, H., Lin, H.: Gender Recognition using Adaboosted Feature. In: 3rd Int. Conf. on Natural Computation (2007)
19. Makinen, E., Raisamo, R.: An Experimental Comparison of Gender Classification Methods. Pattern Recognition Letters 29(10), 1544–1556 (2008)
20. Makinen, E., Raisamo, R.: Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces. IEEE Trans. on PAMI 30(3), 541–547 (2008)
21. Lu, H., Yingjie, H., Yenwei, C., Deli, Y.: Automatic Gender Recognition Based on Pixel-Pattern Based Texture Feature. J. of Real-Time Image Processing, 109–116 (2008)
22. Xu, Z., Lu, L., Shi, P.: A Hybrid Approach to Gender Classification from Face Images. In: 19th Int. Conf. on Pattern Recognition, pp. 1–4 (2008)
23. Ramesha, K., Srikanth, N., Raja, K.B., Venugopal, K.R., Patnaik, L.M.: Advance Biometric Identification on Face, Gender and Age Recognition. In: Int'l Conf. on Advances in Recent Technologies in Communication and Computing, pp. 23–27 (2009)
24. Lu, L., Shi, P.: A Novel Fusion-Based Method for Expression-Invariant Gender Classification. In: Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, pp. 1065–1068 (2009)
25. Xia, H., Hassan, U., Ian, P.: Gender Classification Based on 3D Face Geometry using SVM. In: Int'l Conf. on Cyber Worlds, pp. 114–118 (2009)
26. Ramesha, K., Srikanth, N., Raja, K.B., Venugopal, K.R., Patnaik, L.M.: Feature Extraction Based face Recognition, Gender and Age Classification. Int. J. of Computer Science and Engineering 2(1), 14–23 (2010)
27. Bui, L., Tran, D., Xu, H., Chetty, G.: Face Gender Recognition Based on 2D Principal Component Analysis and Support Vector Machine. In: IEEE 4th Int. Conf. on Network and System Security (2010)

28. Jabid, T., Kabil, H., Chae, O.: Gender Classification using LDP. In: IEEE Int. Conf. on Pattern Recognition, pp. 2162–2165 (2010)
29. Li, Y., Zhang, Y., Zhao, S.: Gender Classification with Support Vector Machines Based on Non-Tensor Prewavelets. In: 2nd Int. Conf. on Computer Research and Development, pp. 770–774 (2010)
30. Rai, P., Khanna, P.: Gender Classification Using Radon and Wavelet Transforms. In: IEEE 5th Int. Conf. on Industrial Information System, pp. 448–451 (2010)
31. Alexandre, L.: Gender Recognition: A Multiscale Decision Fusion Approach. Pattern Recognition Letter 31(1), 1422–1427 (2010)
32. Yan, C.: Face Image Gender Recognition Based on Gabor Transform and SVM. In: Shen, G., Huang, X. (eds.) ECWAC 2011, Part II. CCIS, vol. 144, pp. 420–425. Springer, Heidelberg (2011)
33. Basmaci, E.S., Kaymakcioglu, U., Kurt, Z.: Comparison of Feature Extraction and Feature Selection Approaches to Decide whether a Face Image Belongs to a Male or a Female. In: IEEE 19th Int. Conf. on Signal Processing and Communications Applications, pp. 522–525 (2011)
34. Sonka, M., Hlavac, V.: Digital Image processing and Computer Vision. Cengage India Learning Private limited, New Delhi (2007)

Text Dependent Voice Based Biometric Authentication System Using Spectrum Analysis and Image Acquisition

Somsubhra Gupta[1] and Soutrik Chatterjee[2]

[1] Department of Information Technology, JIS College of Engineering, Kalyani, India
gsomsubhra@gmail.com
[2] Department of Information Technology, Bengal Engineering and Science University, Howrah, India
chatterjee.soutrik@gmail.com

Abstract. Biometrics is concerned with identifying a person based on the physical or behavioral traits of him such as face, fingerprints, voice and iris. With the pronounced need for robust human recognition techniques in critical applications such as secure access control, international border crossing and law enforcement. Biometrics is a viable technology that can be used into large-scale identity management systems. Biometric systems work under the assumption that many of the physical or behavioral traits of humans are distinctive to an individual, and that they can be precisely acquired using sensors and represented in a numerical format that helps in automatic decision-making in the context of authentication. In the presented approach effort has been made to design a Voice based Biometric Authentication system with desired aspiration level.

Keywords: Authentication, Biometric, Image acquisition, Spectrum.

1 Introduction

Nearly 40 years ago, IBM suggested that a computer user could be recognized at a computer terminal "By something he knows or memorizes.... By something he carries... By a personal physical characteristic". This analysis was done in the context of computer data security - remotely recognizing those authorized to access stored data - and specifically referenced voice recognition [1, 2, 3] as a "personal physical characteristic" useful for human recognition.

Recent data on mobile phone users all over the world, the number of telephone landlines in operation, and recent voice over IP (VoIP) networks deployments, confirm that voice is the most accessible biometric trait as no extra acquisition device or transmission system is needed. This fact gives voice an overwhelming advantage over other biometric traits as well as authentication trait [4, 5], especially when remote users or systems are taken into account. Biometric characteristics or traits are often compared under the following criteria:

- Universality: The biometric characteristic should be universally available to everyone.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 61–70.
springerlink.com

- Distinctiveness: The biometric characteristics of different people should be distinctive.
- Permanence: The biometric characteristic should be invariant over a period of time that depends on the applications
- Performance: The biometric authentication system based on the biometric characteristic should be accurate, and its computational cost should be small.
- Acceptability: The result of a biometric authentication system based on certain biometric characteristic should be accepted to all users.
- Circumvention: Biometric characteristics that are vulnerable to malicious attacks are leading to low circumvention.

Considering all these aforementioned criteria, voice serves as quite a handy trait for biometric authentication.

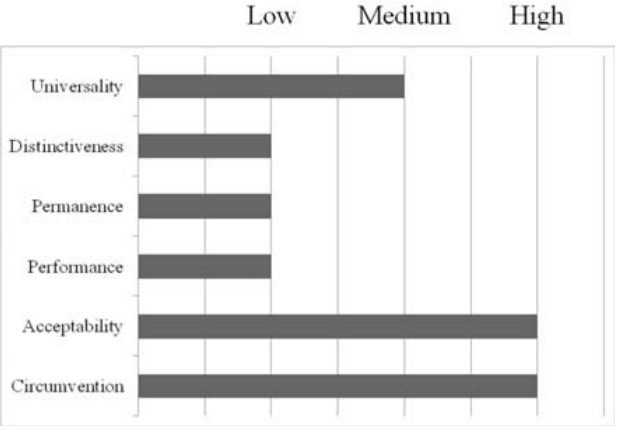


Fig. 1. Voice as a Biometric trait

Different types of speaker recognition technologies [6, 7, 8] with different potential applications can be broadly categorized into two. Firstly, text-dependent technologies in [9], which the user is required to utter a specific key-phrase (such as "password" etc.). The security level of password based systems can then be enhanced by requiring knowledge of the password, and also requiring the true owner of the password to utter it.

The second type of speaker recognition technologies is those known as text-independent [9]. These are the driving factor of the remaining two types of applications, namely speaker detection and forensic speaker recognition. Since the main source of information encoded in the speech is linguistic content, text-independency has been a major challenge and the main subject of research of the speaker recognition community in the last two decades.

Here a voice recognition system of the former kind is taken into consideration. Since the text-dependent systems provide the speaker recognition system with extra information, thus they generally perform better than in the text-independent systems.

Among the existing applications in the study of voice based biometric authentication [10, 11] the prominent one was designed by Purdy Ho and John Armington [12]. Their Dual-Factor Authentication System consists of speaker verification and token technology. This is a voice authentication system combining speaker verification and token technology. This dual-factor system is designed to prohibit imposture by pre-recorded speech followed by text-to-speech voice cloning (TTSVC) technology and furthermore to rationalize the inconsistencies of audio characteristics of various devices. The token device generates and prompts a onetime passcode (OTP) to the user. The spoken passcode is then sent simultaneously to a speaker verification module in order to verify user's voice, and to a speech recognizing module, which converts the passcode to text and validates it. Thus, the passcode protects against recorded speech or voice cloning attacks and speaker verification module defends against the use of a stolen or lost token device.

Some inventions are concerned with a method for voice authentication on a device. The method can include gathering one or more spoken utterances from user, then identifying a phrase corresponding to the spoken utterances and recognizing a voice print from them, determining a specific device identifier associated with the device. The purpose is to authenticate the user based on the phrase, the device identifier and the voice print. A variety of spoken utterances can be used for creating the voice print. The voice print is a vocal tract configuration that is unique to a vocal tract of the user. Depending on the authentication, access can be granted to one or more resources that have a communication with that particular device.

2 Methodological Aspects

The methodological aspects of the proposed work have been discussed in the following sections:

2.1 Identifying Information in Speech Signal

Speech production is a extremely complex process whose result depends on many variables at different levels, including from sociolinguistic factors (e.g. level of education, linguistic context and dialectal differences) to physiological issues (e.g. vocal tract length, shape and tissues and the dynamic configuration of the articulatory organs). These multiple influences will be simultaneously present in each speech act and some or all of them will contain specificities of the speaker.

For that reason, different levels and sources of speaker information are needed to be clearly distinguished and clarified so as to extract requisite information in order to model speaker individualities.

2.2 Multiple Levels of Information

Experiments with human listeners have shown that humans recognize speakers by a combination of different information levels, and what is specially important, with different weights for different speakers (e.g. one speaker can show very characteristic pitch contours, and another one can have a strong nasalization which make them "sound" different).

Automatic systems will intend to take advantage of the different sources of information available, combining them in the best possible way for every speaker.

Idiolectal
Idiolectal characteristics of a speaker are at the highest level that is usually taken into account by the technology to date, and describe how a speaker uses a specific linguistic system. This "use" is determined by a multitude of factors, some of them quite stable in adults such as level of education, sociological and family conditions and place of origin. But there are also some high-level factors which are highly dependent on the environment, as e.g., a male teacher does not use language in the same way when talking with his colleagues at the school (sociolects), with his family at home, or with his friends' playing cards.

Prosody
Prosody is the combination of instantaneous energy, intonation, speech rate and unit durations, make provisions for natural, full and emotional speech. It determines prosodic objectives at the phrase and discourse level, and define actions to comply with those objectives.

Short-term spectral characteristics:
Finally, at the lower level, we find the short-term *spectral* characteristics of the speech signals, directly related to the individual articulatory actions related with each phone being produced and also to the individual physiological configuration of the speech production apparatus. This spectral information has been the main source of individuality in speech used in actual applications.

Spectral information intends to extract the peculiarities of speaker's vocal tracts and their respective articulation dynamics. Two types of low level information has been typically used, *static* information related to each analysis frame and *dynamic* information related to how this information evolves in adjacent frames, taking into account the strongly speaker-dependent phenomenon of co-articulation, the process by which an individual dynamically moves from one articulation position to the next one.

3 Procedural Aspect

The model has been designed with the consideration of the following facts:

3.1 Collection of Voice Samples

Voice samples of a particular user are collected to construct an optimally exhaustive voice-base i.e. the collection of voiceprints of a particular user. Procedure involving collection of voiceprint samples are presented in the following:

- The samples are collected with the help of voice recorder utility of default Operating System.
- The user voice is captured by microphone.

- Various voice samples of the same person uttering the same word are collected under different ambiences so as to minimize the external factors.
- It is known that no two successive utterances of the same word of a particular human being are exactly the same. Different samples are collected speaking in different manners so as to minimize the effect imposed by it.

3.2 Spectrum Analysis

This phase deals with the analysis of the characteristics of spectrum [13, 14, 15] to identify the properties of its source. It may consist of,

- The distribution of a trait of a physical system, especially:
- The distribution of energy emitted by a source.
- A graphic representation of such a distribution.
- A range of values of a quantity or set of quantities related to that.

4 Illustrative Case Studies

The software **Sonic Foundry Sound Forge 6.0** for spectrum analysis has been used for testing purposes and data assistance. In the Spectrum analysis graph of the software Spectrum analysis is performed on a particular range of frequency.

The frequency range is specified as 20 Hz to 22,000 Hz in order to analyze the human voice. This matches the frequency range of human auditory system. The amplitude of the power spectrum at each point of the frequency is displayed in the Spectrum analysis graph within that frequency range.

The Spectrum analysis graphs of the utterances of the word "soutrik" by a particular user are shown at the next page.

Minimization of ambience noise of the collected voice samples have been performed under different moods with practical significance. The results obtained there in are presented from Figure 2(a) to 2(c) in the following:

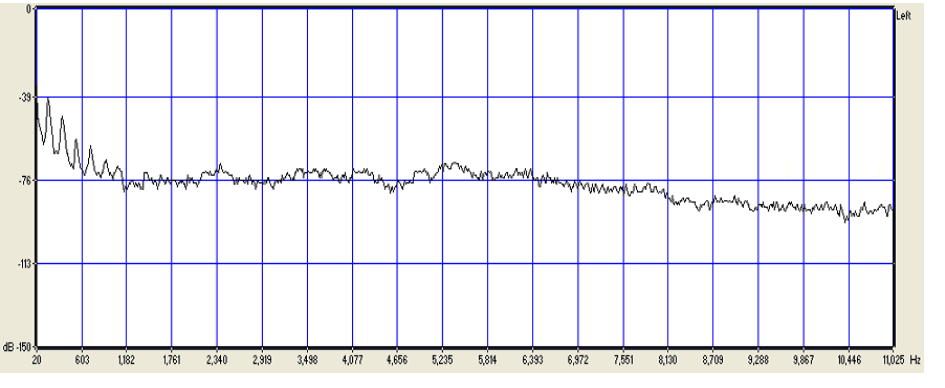


Fig. 2(a). Voice sample – Soutrik under Mood Indexed 1 (Soutrik)

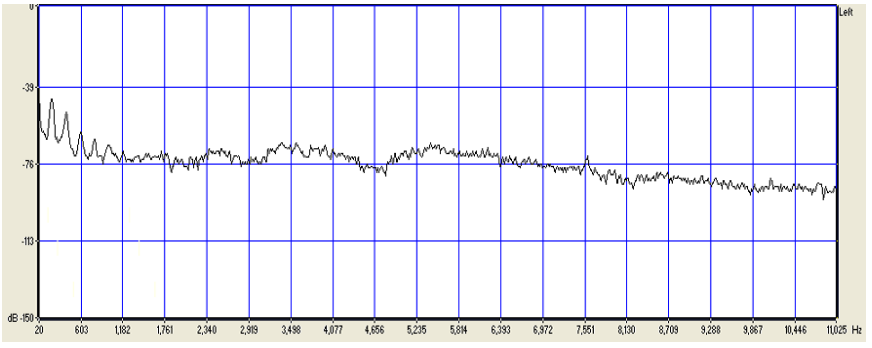


Fig. 2(b). Voice sample – Soutrik under Mood Indexed 2 (Soutrik)

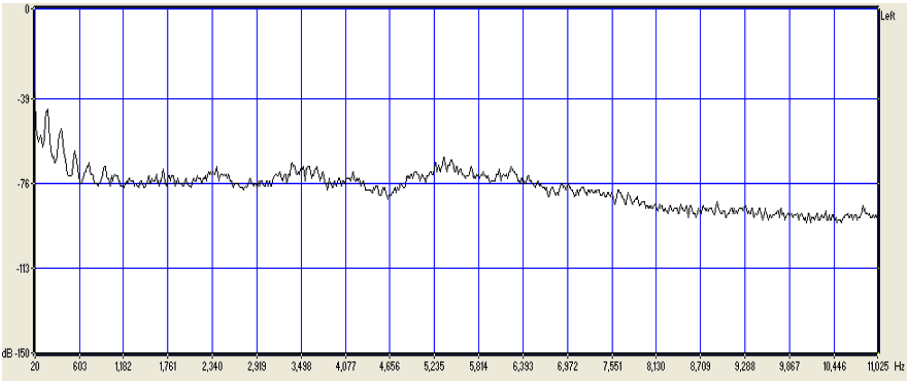


Fig. 2(c). Voice sample – Soutrik under Mood Indexed 3 (Soutrik)

4.1 Construction of Rule Base

A rule base to authenticate a user based upon the features of their voice has been constructed. The construction of rule base includes two steps:

1. Extraction of features of voice of that particular person. The following features from the Spectrum analysis graph are extracted:
   - Pitch of voice of a particular person,
   - Amplitude of voice at a particular frequency of voice for a particular person.
2. Finding out some patterns or trends among the extracted features from the Spectrum analysis graph to construct some rules to uniquely identify that person.
   By analyzing the data extracted from the Spectrum analysis graph, some interesting patterns and trends corresponding to individual are found.
   These patterns or trends are generally cluster wise independent i.e. for a particular cluster of frequency someone may observe a particular pattern or trend that is prevalent in all the voice samples of a particular person taken, whatever the ambience or speaking style may be. From these patterns some production rules are constructed (depending upon the cluster of frequency under examination) to uniquely identify a person's voice.

The change of voice of a person with respect to time is also taken into consideration by making the rule base dynamic, that is, the changing features of a person's voice is taken into account by modifying the rule base according to the newly acquired features of his/her voice.

4.2 Steps in Authentication Mechanism

Based on those samples, Spectrum Analysis has been performed.

- Patterns common to a particular person whose voiceprint to be stored in the knowledgebase are found. The patterns are recorded so as to authenticate individuals
- Based on those patterns which are different to others, it becomes easy to restrict unauthorized access.
- The graphical representations of the above experiments are shown in the next.
- After that, screenshots of an authorized and unauthorized user are shown.

5 Case Results

Graphical representation of voiceprint of an authorized person in comparison with others unauthorized person are presented in the following Figures 3(a) to 3(c):

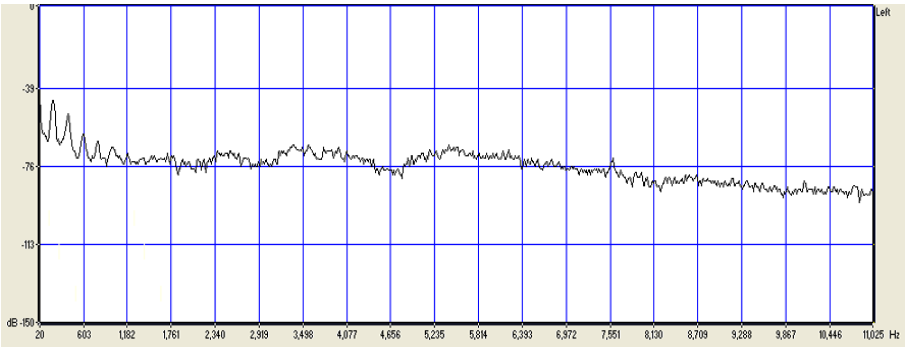


Fig. 3(a). Voice sample _soutrik 3 (authorised)

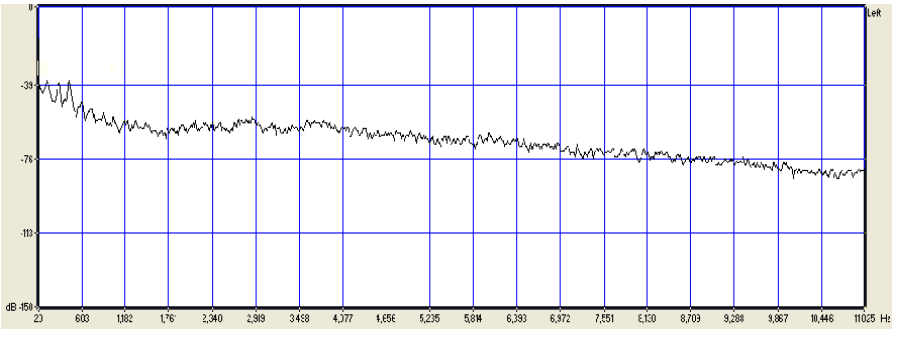


Fig. 3(b). Voice sample _rahul (unauthorised)

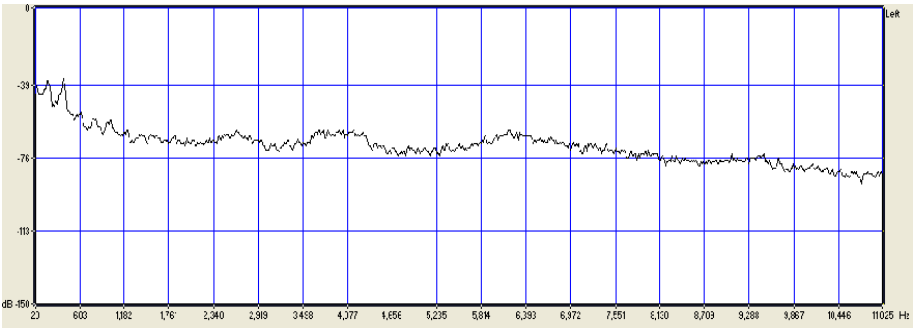


Fig. 3(c). Voice sample _ssg sir (unauthorised)

Obtained results during testing trials are presented in the screenshots Fig.4 and Fig. 5:

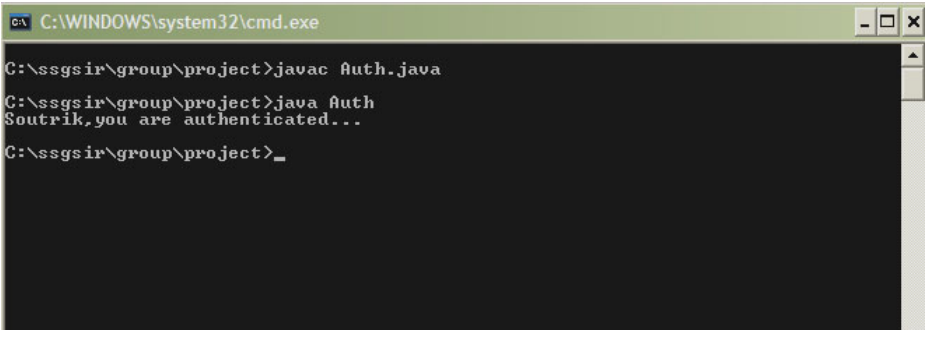


Fig. 4. Voice sample _soutrik (authorised)

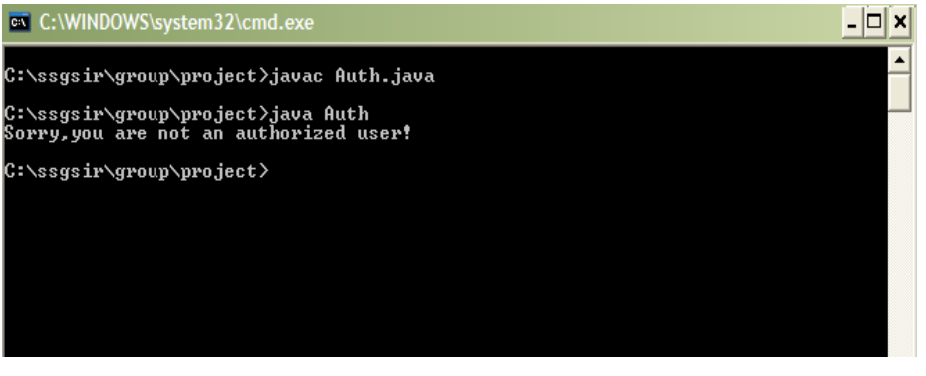


Fig. 5. Voice sample _sutapa (unauthorised)

6 Benchmarks and Databases

The first databases used for text-dependent speaker verification were databases not specifically designed for this task like the TI-DIGITS and TIMIT databases. One of the first databases specifically designed for text-dependent speaker recognition research is YOHO. It consists of 96 utterances for enrollment collected in 4 different sessions and 40 utterances for test collected on 10 sessions for each of a total of 138 speakers.

However, the YOHO database has several limitations. For instance, it only contains speech recorded on a single microphone in a quiet environment and was not designed to simulate informed forgeries (i.e. impostors uttering the password of an user). More recently the MIT Mobile Device Speaker Verification Corpus has been designed to allow research on text-dependent speaker verification on realistic noisy conditions.

The difficulty in comparing different text-dependent speaker verification systems is that these systems tend to become language dependent. Consequently researchers tend to present their results in their custom database in which it is almost impossible to make direct comparisons. The comparison of different commercial systems is even more difficult.

7 Conclusions

Voice biometric authentication has a number of advantages over other biometric technologies, and will continue to grow in popularity due to the ease of use, user acceptance, and remote authentication capabilities. Its enhance impact and potential as authentication process is well investigated and well documented [16, 17]. The future uses are only limited by the scope of our imagination.

The limitations of the proposed approach can be stated as for efficient functioning, the system requires a vast knowledgebase. Moreover the authentication mechanism shows a high degree of configurational dependency.

The proposed work can be further enhanced so as to make this system platform-independent and adaptive [18, 19]. Incorporation of low-level programming in order to authenticate during booting time can be a potential requirement in the future scope of study.

References

1. Ramachandran, R.P., Zilovic, M.S., Mammone, R.J.: A Comparative Study of Robust Linear Predictive Analysis Methods with Applications to Speaker Identification. IEEE Transactions on Speech and Audio Processing 3(2), 117–125 (1995)
2. Atal, B.S.: Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. J. Acoust. Soc. Amer. 55, 1304–1312 (1974)
3. Rabiner, L.R., Schafer, R.W.: Digital Processing of Speech Signals. Prentice-Hall, Englewood Cliffs (1978)

4. Gupta, S., Bhattachryya, M., Chatterjee, S.: Variable length cumulative key size and seed based symmetric key cryptographic algorithm using composite 3-D transposition substitution and Chaining Technique. In: Proc. of the IEEE International Conference on Advances in Communication, Network and Computing (CNC) 2010, pp. 109–113 (2010)
5. Gupta, S., Sengupta, S., Bhattachryya, M., Chatterjee, S., Sen Sharma, B.: Cellular Phone Based Web Authentication System Using 3-D Encryption Technique under Stochastic Framework. In: Proc. of the 1st IEEE Asian Himalayas International Conference on Internet (AH-ICI) 2009, pp. 1–5 (2009)
6. Doddington, G.R.: Speaker recognition—Identifying people by their voices. Proc. of IEEE 73, 1651–1664 (1985)
7. Lee, C.-H.: On robust linear prediction of speech. IEEE Trans. Acoust., Speech, Signal Processing 36, 642–650 (1988)
8. Ma, C., Kamp, Y., Willems, L.F.: Robust signal selection for linear prediction analysis of voiced speech. Speech Commun. 12(2), 69–81 (1993)
9. Rosenberg, A.E., Soong, F.K.: Evaluation of a vector quantization talker recognition system in text independent and text dependent modes. Comp. Speech Language 22, 143–157 (1987)
10. Furui, S.: Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust., Speech Signal Processing ASSP-29, 254–272 (1981)
11. Assaleh, K.T., Mammone, R.J.: New LP-derived features for speaker identification. IEEE Trans. Speech Audio Processing 2, 630–638 (1994)
12. Ho, P., Armington, J.: A Dual-Factor Authentication System Featuring Speaker Verification and Token Technology. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688, Springer, Heidelberg (2003)
13. Rothacker, R.J., Mammone, R.J., Davidovici, S.: Spectrum enhancement using linear programming. In: IEEE Int. Cunf. Acoust. Speech Signal Processing, Tokyo, Japan, pp. 43.10.1–43.10.4 (1986)
14. Therrien, C.W.: Discrete Random Signals and Statistical Signal Processing. Prentice-Hall, Englewood Cliffs (1992)
15. Rabiner, L.R., Juang, B.-H.: Fundamentals of Speech Recognition. Prentice-Hall, Englewood Cliffs (1993)
16. Duda, R.O., Hart, P.E.: Pattern Classification and Scene Analysis. Wiley, New York (1973)
17. Oppenheim, A.V., Schafer, R.W.: Discrete-Time Signal Processing. Prentice-Hall, Englewood Cliffs (1989)
18. Farrell, K.R., Mammone, R.J., Assaleh, K.T.: Speaker recognition using neural networks versus conventional classifiers. IEEE Trans. Speech Audio Processing 2, 194–205 (1994)
19. Zilovic, M.S., Ramachandran, R.P., Mammone, R.J.: A fast algorithm for finding the adaptive component weighted cepstrum for speaker recognition. IEEE Trans. Speech Audio Processing 5, 84–86 (1997)

Interactive Investigation Support System Design with Data Mining Extension

Somsubhra Gupta[1], Saikat Mazumder[2], and Sourav Mondal[3]

[1] Department of Information Technology, JIS College of Engineering, Kalyani, India
gsomsubhra@gmail.com
[2] Department of Computer Application, JIS College of Engineering, Kalyani, India
saikatmaz@gmail.com
[3] Siemens Information Systems Ltd, Kolkata, India
sourav85mondal@yahoo.com

Abstract. The Government and non Government investigation organizations (e.g. CID, CBI etc) are equipped with huge framework of databases constantly being operated and analyzed by high-end professional officials for updating and retrieving facts which assists these organizations in information requirement for investigation, investigation proceedings and finally for solving the case. This Interactive Investigation Support System is designed for the purpose of supporting crime investigation conducted under the jurisdiction of local Police Station (including District level authorization) and related issues of administrative bureaucratic hierarchy. Within the scope of this support system, fields of crime investigation have been streamlined to crimes based on vehicle theft. There is an option for providing support to search among old criminal(s) who has already committed similar crime on the same area using data mining technique. It will not detect the criminal(s). It only gives an additional support for decision making. The objective of the system is to encompass these aforesaid dynamic features operating within a large framework of databases. It interacts with a usability tested Graphical User Interface, provides user-friendly searching method – which computationally less complex, interactive and with least bug.

Keywords: Crime investigation, Data Mining, FIR report, Preprocessing, Support Systems.

1 Introduction

The reputed and high end professional in Government investigation organizations (like CID, CBI, RAW etc) uses sophisticated computerized techniques for storing, updating and searching - suspect, criminal and crime related records [1, 2]. But, this system has been proposed and software has been developed to supplement with the purpose that the local Police Station and their higher authority will have a convenient, user-friendly and secure system which updates any suspect / criminal's record with judicial out come of related cases. It will cut the time spend to search any crime records manually, helps inter-police station information sharing, search an

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 71–81.
springerlink.com

information in user friendly but robust technique, it merges judicial and police station records thus provide a combined platform for these two main administrative pillar of the society.

It is to be mentioned in the context of the project that the issues on Crime investigation is not widely circulated in the literature [3, 4]. During the course of the work and there after this project has experienced lack of literature support as most of the available works are descriptive in nature and don't incorporate detailed structure. It is an existing perception that researches on crime investigation to the extent of defence research are confidential issues. However, the present work is aimed at assisting the existing framework purely from the academic viewpoint.

1.1 Salient Features of the Proposed Work

The proposed system incorporates the following features:

- The system possesses defined security measures, because it handles secure data. Entry to the system is password protected and the username will decide the privileges of the user for the system. The log in details of a user has been stored as well as his crucial course of actions also, by firing trigger.
- Facility to store in detail information of a case including FIR details, names of IO with case proceeds, suspect person's details, informant/complaint details, theft or seized materials details, arrest warrant details, arrested person(s)'s details, Final report details, case proceedings details and judicial out come of the case including judgment of the person(s) who is/are involved in that particular case.
- Easy to use searching tool, which is made such as a user can search any number of fields he wants. Searching can be done with three main categories: case based search, persons based search and search based on property stolen/seized (properties including automobile, cultural property, general property and fire arms).
- The system is connected via LAN to a central database server. Every Police Station stores the information into that database and this is shared by all.
- Data Mining is used to give a suggestion in the current case (data mining is streamlined to vehicle thieving only).
- Avoided repetition of details about a particular criminal/suspect; when a new entry is going to happen, system will check if old records of similar type exist or not.
- The system updates every criminal's police records automatically with the judicial out come of a case related to him. The system manages information of two main part of the administration: Police System and Judicial system; more over it also manages detail information of a criminal (physical, social and crime records) as well as his photograph and finger print.
- Only administrative privilege gives user the right to create new user id and change password of existent user and delete record of a particular case. Deleting record is made such that if a FIR record is deleted, the corresponding arrest details, Final report and Judicial record also deleted for that case. But if judicial record is deleted for a case FIR record and arrest details for that case will remain.

1.2 Steps in the Case Proceedings

The designed system is flexible enough to work as step by step entry method as well as it can take random entry for each step (provided it doesn't conflict with the security and concurrency of the database). Though, there is a general step by step entry and proceedings mechanism adopted in the system from actual criminal proceedings [2, 3] of administrative hierarchy.

Step 1: First Information Report
This step stores information of a FIR including date, time, place of offence and FIR content. Three pop up windows are there to store information about informant/complaint, suspect and the stolen property details.

Step 2: Property Stolen Information
Here, the user can store full details of stolen property if he doesn't entry it at the time of FIR has been lodged. The conditions are, only recorded FIR no's of a Police Station's for a particular date are available for entry.

Step 3: Arrest / Surrender Information
Here, two ways entry method can be possible viz. This is either a recorded criminal in past case(s) or a new suspect who is yet to be convicted. If there is not any record exist of this suspect, system will prompt to save his details (physical and social); if old criminal system will conclude that his records all ready exist. Old cases (if exist) of this particular criminal also been shown by the system. User can decide which past case(s) should be related with this present case.

Step 4: Probable search of criminal
A search from listed criminal of same type of crime can be done. Data mining is used for this purpose. Here, the software takes the case no as input then it matches all the previous record to find same type of cases. When it finds the similar cases, it collects the information about the arrested/punished criminals of that case. This information is compared with the suspect details of the current case. Then a comparative analysis of suspect and past criminals has been done.

Step 5: Final Report: Final report detail for a particular suspect in a particular case is stored. Here information has been stored such as type of final report (i.e. whether it's a charge sheet), whether charge sheet is original or supplementary, and status of accused i.e. whether he is in judicial custody or bailed out.

Step 6: Court Disposal: The judicial out come of a case and punishment given by the court are stored. The crime record of the criminal involved in this case has been updated with the judicial feedback as well as the case status for the police enquiry has been updated with the case status (pending, reviewed or closed).

2 System Planning

The system has three-level architecture. These levels according to their hierarchical authentication status are, namely- the territorial level, the zonal level and the local level.

The core part of the system is Territorial level which holds the Administrative power and privileges of the total system. This level of police administration (viz. Superintended of Police) can access and interfere with the entire database and gives the permission to act upon a certain situation to the zonal level. At this level only issues of foremost priority like inter-district matters and state related matters are dealt with. The next layer is Zonal level (viz. SDPO) which is predominantly the most active level under the designed system and at this layer, interference by the authority is done in the fields of intra-police station matters, maintenance of sub-judice register and issues regarding inter-related/co-related criminal activities [2, 5]. Important and final decision making ascendancy rests on this level (except for a few for which it will depend on its top hierarchy).

Zonal level needs permission to take some decision from the supervisor, i.e., from the Territorial level, but on the most of the time it has the control power to handle the major number of cases. At the base of the hierarchical layer is the Local level. This part of the designed system is privileged to work within the periphery of the local Police Stations. This level is entirely the operational level as all the preliminary activities (like FIR entry, updating criminal related information or investigation report). But almost all the decision required in this level is to be authenticated through its immediate superior level i.e. the Zonal level. They always need permissions from Zonal level, to take decision about the cases where more than one local level is interacting. Though Territorial level holds administrative privilege in the system, but maximum decision making and control power involved in the Zonal level. The three levels will be connected through intra-net.

The following is the access privileges designed to grant to various levels according to the specification and practice.

Local level:

- Entry of FIR (Contain FIR number, date, IPC sections, gist etc).
- View, search and print FIR related details.
- Submit and update Interrogation Report.
- Update of the crime details of a criminal (like STATUS IN A CASE- (arrested, accused, charge-sheet given etc), aliases, gang).
- Update the detailed information of a criminal as individual and as a gang member (Description, Identifying Particulars, and associates/co-accused).

Zonal level:

- Entry and update of Police Station information.
- Entry and update of individual details of a Police Man.
- The case details (handled by whom, case start date, involved police stations)
- Connecting the related crimes, second degree association.
- Interfere in the case(s) where more than one Police Station is involved.
- Keep track of judicial matters - subjudice register, trial court, dates of case(s), charge-framing, court dates etc.
- View, modify and print the above sayed related report(s).

Territorial level:

It has the Administrative Privilege. All access granted. Except intervention in the matter of major acts and inter-district matters and state related issues.

Can view, search or take report on the various issues which are not directly linked with main functionality of the Investigation Support System, e.g. Nil-arrest cases, various reports including, Inter-P.S. criminals,

2.1 Control Flow

The control flow of the proposed system is presented in the following diagram (Fig. 1).

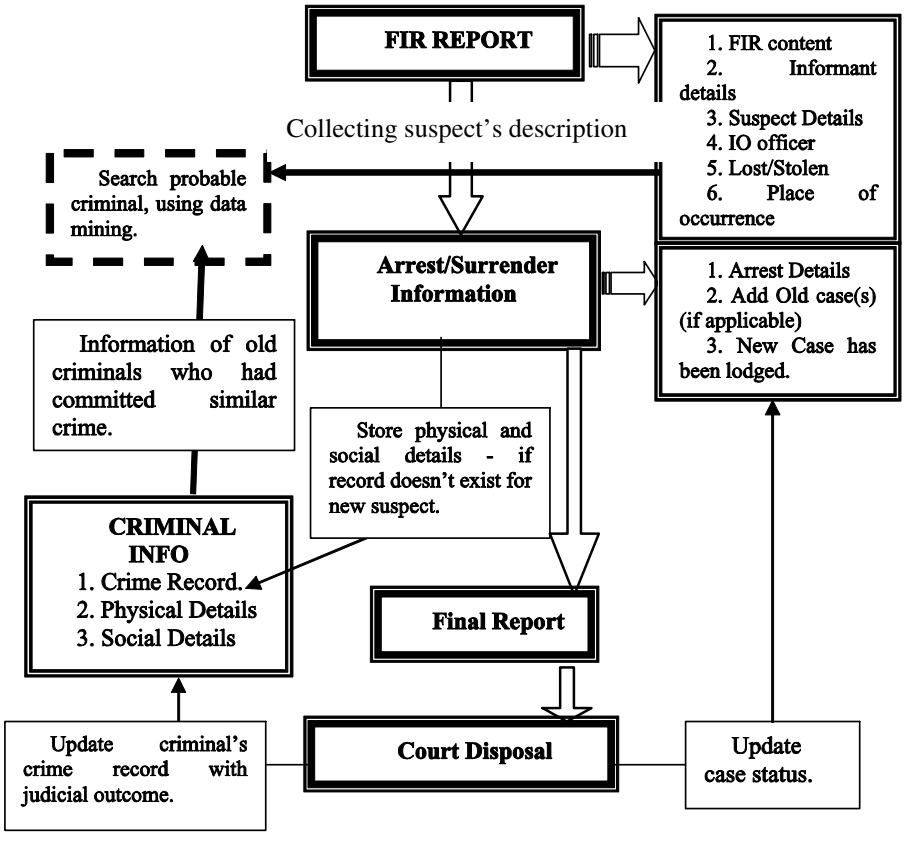


Fig. 1. Work-flow of the system control mechanism clearly indicating the methodological aspects

3 Use of Data Mining Technique

Before implementing the technique to determine probable criminal [6], the concepts of data mining [7, 8, 9] is used. The steps involved in it are follows:

1. Selection

This stage is concerned with selecting or segmenting the data that are relevant to some criteria, in the context of this support system. It's the set of rules that will be derived from analyzing the vehicle theft crime records (For both Two Wheeler, and Four Wheeler or Other Vehicle) of the past three years in the districts of North 24 Pargana, for example, the type of vehicle being stolen and the related places of theft or the related place of recovery of that particular type of vehicle as well as to maintain a track from where usually the involved person's in the crime are related.

2. Preprocessing

Preprocessing is the data cleaning stage where unnecessary information is removed, like when considering the theft cases of cars it is unnecessary to keep the track of information about the engine no and chassis no of each vehicle to find a particular pattern in theft. In the context of this investigating support system the database is strongly bound and the data taken and stored are filtered in the time of searching probable and only valuable data are taken into account for searching.

3. Transformation

The vehicle theft related data is transformed in order to be more suitable for the task of data mining. In this stage the data is made usable and navigable.

4. Knowledge Discovery

This stage is concerned with the extraction of patterns from the data. The required integrated data are evaluated by matching with the knowledge base which is formed from the integrated facts picked out of the bounded database on the basis of set of pre-defined rules. For instance, a rule will exist that will describe the relation between the type of automobiles and the location of the group of the criminals who steals them, and then one can deduce the pattern of theft of the particular type of automobile linked with that particular location, and a set of rule can be made.

5. Interpretation and Evaluation

The evaluated patterns obtained in the data mining stage are converted into knowledge which is used for decision making, as after the pattern evaluation of the afore mentioned relation, the necessary features of the probable suspect can be obtained in case of the next property theft which is the ultimate decision making.

3.1 Search Control Flow in the State Space

State space represents the collection of all feasible records available in the database along with the probable operations to be carried out in a knowledge driven way under

the framework of Data Mining. The control flow for searching probable criminal is presented in the following Figure 2.

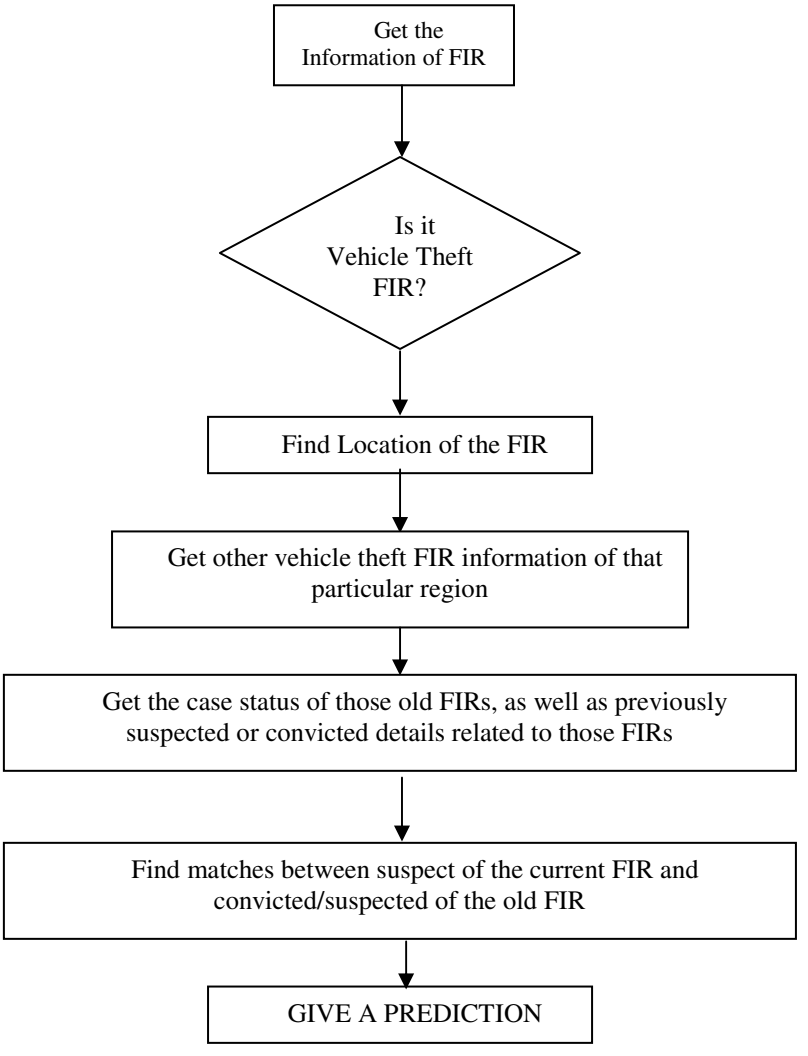


Fig. 2. Searching a probable criminal

3.2 Entity Relationship Mode

The entity relationship model of the proposed system is presented in the following:

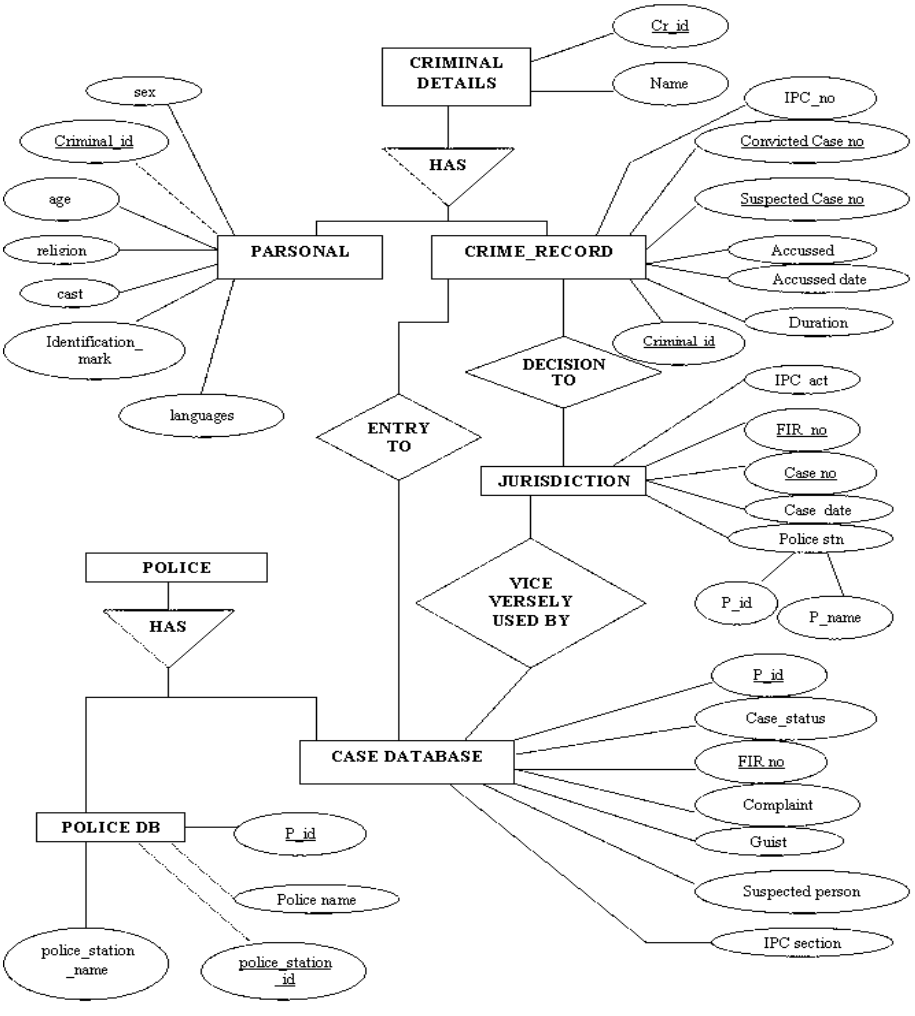


Fig. 3. ER model of the proposed systems

3.3 Data Flow Diagram

The DFD at level 1 is presented in the following figure 4:

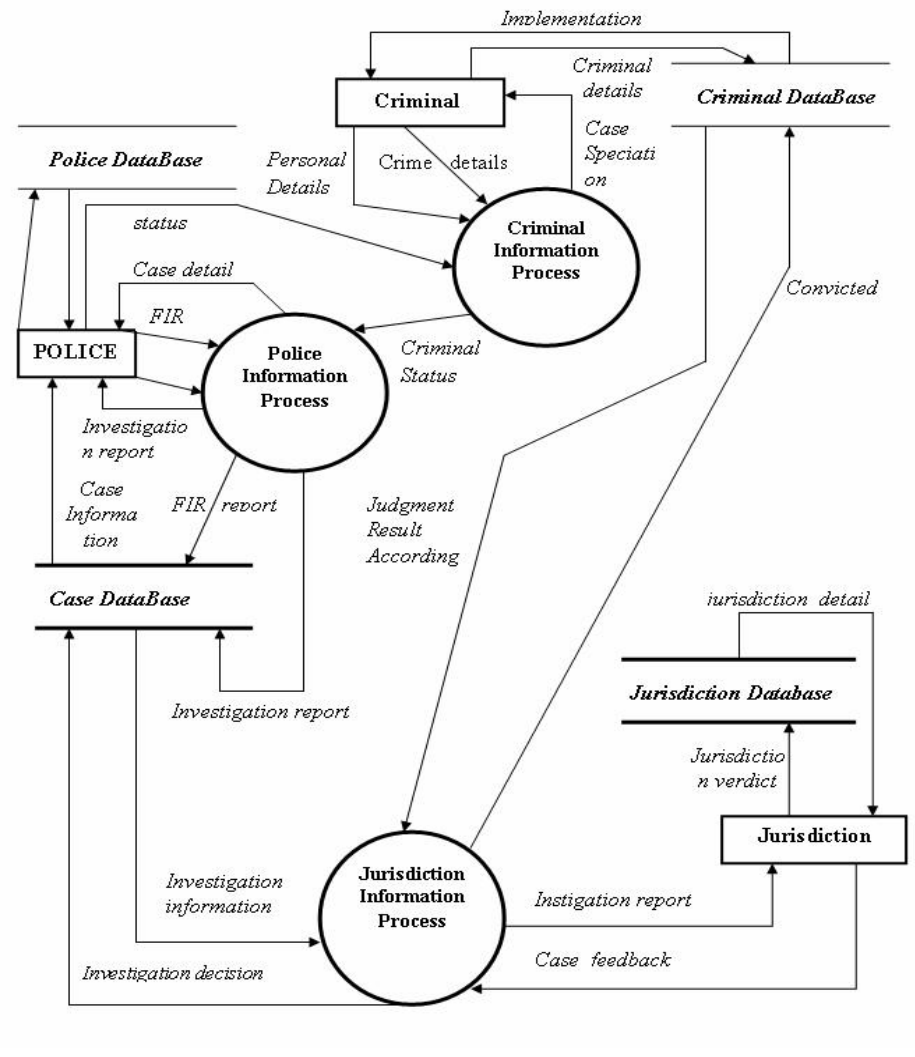


Fig. 4. DFD at Level 1

4 Limitations and Constraint

The system is stringent in some aspects because of the security reason. When arrested details, Final Report or Court Disposal is stored for an particular case no, the system will force the user to chose the case no from system defined instead of some input

given by user. As for a given case multiple no of arrest details, Final Report or Court Disposal report is possible, thus system will take multiple numbers of times these records. If any user put multiple records by mistake system can't check it.

There is not any provision to edit the information stored in the system, because editing any past record is very much crucial for the proceedings of a case. Such as, if name of a suspect or the physical description of a suspect is changed then it will differ hugely with the arrested details and the judicial out come stored in the system. Thus, the only provision is kept in the system is deleting the record. If wrong value is inputted, then that entire record must be deleted then it should be re-enter.

The software which is developed to keep information about criminals and crime cases and help officers in their investigation is quite available to the high end professional in investigative organizations. But, the aim of this software is to help the police system. Thus, their course of action towards a case is to be followed as well as information is to be collected about the judicial proceedings of a case. Now, both police authority and judicial system keep the data and information about criminal proceedings strictly confidential. To develop this system successfully, two things need to know for sure: firstly, the technique followed by both police authority and judicial system to store their data in a synchronized order and their interpretation to those data. Secondly, lots of data needed for a particular type of crime to study the data. To implement data mining technique in a crime pattern, it is the primary need that enough amount data should be present for mining.

Both of these issues are confidential matter for police authority. Thus the study of their work pattern mainly done by consulting experienced officers. The authority is agreed to give small amount of real data to study for implementing Data Mining. But, again, the amount of data is small enough to successfully implement it in software.

5 Conclusions

In the proposed work effort has been made to extend the conventional system methodologies to develop a interactive investigation system. This has been proposed and supplemented via designed software under MS.NET Framework. The main focus is on identifying interesting patterns using data mining in order to find crime trend in a specific domain. This production system of which has been proposed by employing association rules on large crime database.

After the completion of the framework, design, search mechanism, report generation and implementing data mining extension, then comes the period of improvisation. Improvisation is an ongoing phenomenon and improvisation craves the path towards improvement. In the presented work, the following extension to the existing features are planned to be embedded.

Statistical trend analysis is to be done in time series in the considered field of crime and thus determining the probable suspects who are more prone to be the criminals for a particular case.

Identification of appropriate heuristic for making the searching procedure faster by working on type representation format and partial matching is a need for enhancement.

Image processing mechanism may be embedded so as to find matches between finger print between existing criminals and thus provide valuable information to the investigation.

References

1. Fayyad, U.M., Uthurusamy, R.: Evolving data mining into solutions for insights. Communications of the ACM 45(8), 28–31 (2002)
2. Schneider, S.: Predicting crime: a review of the research. Department of Justice Canada, pp. 1–2 (2002)
3. Chau, M., Xu, J., Chen, H.: Extracting meaningful entities from police narrative reports. In: Proceedings of the National Conference for Digital Government Research (dg.o 2002), Los Angeles, California, USA (2002)
4. Hauck, R.V., Atabakhsh, H., Ongvasith, P., Gupta, H., Chen, H.: Using Coplink to analyze criminal-justice data. IEEE Computer 35(3), 30–37 (2002)
5. Wang, G., Chen, H., Atabakhsh, H.: Automatically detecting deceptive criminal identities. Communications of the ACM (accepted for publication, forthcoming)
6. Chen, H., Lynch, K.J.: Automatic construction of networks of concepts characterizing document databases. IEEE Transactions on Systems, Man, and Cybernetics 22(5), 885–902 (1992)
7. Malathi, A., Santhosh Baboo, S., Anbarasi, A.: An intelligent Analysis of a City Crime Data Using Data Mining. In: Porc. of International Conference on Information and Electronics Engineering IPCSIT, vol. 6, pp. 130–134. IACSIT Press, Singapore (2011)
8. de Vel, O., Anderson, A., Corney, M., Mohay, G.: Mining E-mail Content for Author Identification Forensics. SIGMOD Record 30(4), 55–64 (2001)
9. Tolle, K.M., Chen, H.: Comparing noun phrasing techniques for use with medical digital library tools. Journal of the American Society for Information Science 51(4), 352–370 (2000)

Pattern Recognition Approaches to Japanese Character Recognition

Soumendu Das and Sreeparna Banerjee

School of Engineering Technology
West Bengal University of Technology, India
sdphotoes@gmail.com, sreeparnab@hotmail.com

Abstract. Optical character recognition is a crucial step in the document retrieval and analysis. However this process could be error prone, especially in Japanese language, where the text is composed from over 3000 characters which can be classified as syllabic characters, or Kana, and ideographic characters, called Kanji. Moreover, Japanese text does not have delimiters like spaces, separating different words. Also, the fact that several characters could be homomorphic, i.e. having similar shape definition could add to the complexity of the recognition process. In the note, a survey has been conducted of some of the approaches that have been attempted to address these issues and devise schemes for Japanese character recognition in texts. Also, our efforts to extract Japanese text using image processing techniques have been described and some of the results have been presented.

Keywords: Japanese character recognition hiragana, katakana and kanji, document retrieval.

1 Introduction

In the present era, one of the major tasks for offices is to retrieve desired documents from huge document databases. In order to convert proper documents to electronic documents, optical character readers are required, but these are error prone. Hence there exists a necessity of efficiently combining optical character recognition along with document retrieval techniques. Document retrieval technique in Japanese is further complicated by the fact that the text comprises of both the syllabic / phonetic characters (Kana) as well as the ideographic characters (Kanji) and similar shape definition of several Japanese characters. Furthermore, Japanese text is not separated by delimiters such as spaces. Homomorphism or similar shape definition for different Japanese characters also poses problems especially in sans serif fonts. This survey describes several approaches described in the literature, for Japanese text recognition and retrieval.

The next section discusses the Japanese language model briefly. Succeeding selections describe the various approaches. A comparison with Korean text retrieval is also included, followed by concluding remarks.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 83–92.
springerlink.com

2 Japanese Language Model

Two aspects of Japanese language, namely, text and scripts, are described below.

2.1 Japanese Text

Japanese text is written using more than 3000 characters, many of which have complex and similar shapes, and the text is not separated by delimiters such as spaces.

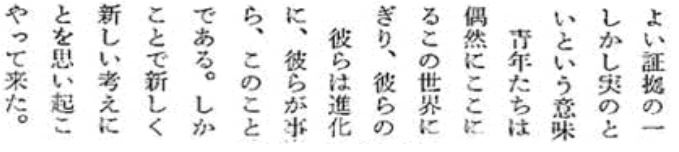


Fig. 1. Sample Japanese Text

2.2 Japanese Text

Japanese writing system has three different characters sets, namely, Hiragana, Katakana and Kanji. For Japanese words, Hiragana (see Fig2a) is used, mostly for grammatical morphemes. Katakana (see Fig2b) is used for transcribing foreign words, mostly western, borrowing and non-standard areas. In addition, diacritic signs like dakuten and handakuten are used (see Fig3 and 4).

- Dakuten are used for syllables with a voiced consonant phoneme. The dakuten glyph (゛) resembles a quotation mark and is directly attached to a character (Foljanty 1984).

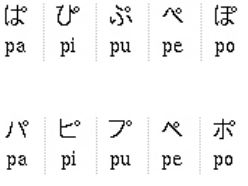


Fig. 2. (a)Hiragana & (b)Katakana **Fig. 3.** Dakuten and **Fig. 4.** Handakuten Alphabets

- Handakuten are used for syllables with a /p/ morpheme. The glyph for a 'maru' is a little circle (゜) that is directly attached to a character (Foljanty 1984).

Kanji are content bearing morphemes. In Japanese text Kanji are written according to building principles like Pictograms (graphically simplified images of real artifacts), ideograms (combinations of two or more pictographically characters) and phonograms (combinations of two Kanji characters).

3 Earlier Attempts

Three approaches have been attempted to overcome the problems due to imperfect recognition, as described below.

3.1 Error Correction

Recognition output is made clear by the error conversion. This can be done using a spelling checker which is capable of integrating characteristic patterns of recognition errors which differ from normal typing errors. A second method is to use linguistic knowledge [1] which includes knowledge about the content of documents [2] in addition to syntactic and lexical knowledge. A third method is the category utilizes vocabulary derived from similar documents in order to improve the word recognition rate [3].

3.2 Document Search without Optical Character Recognition

The document processing system "Transmedia Machine" is used for this purpose [4]. Character images of scanned documents are encoded into two binary features for each character succeeded by a "string matching" based on incomplete codes. Word-level encoding [5] has been proposed as a more reliable alternative. Searching for text passages in document image database and subsequent pattern matching using a number of feature descriptors has also pattern been proposed [6].

3.3 Error Tolerant Search System

In order to make the search system tolerant of recognition error, multiple candidates have been used in the search process [7]. The optical character recognition keeps multiple candidates for ambiguous recognition and outputs them as a result text. Segmentation ambiguities [8] can also be included, with multiple hypotheses in both character segmentation and recognition represented as a network of hypotheses.

4 Document Retrieval Tolerating Character Recognition Errors

Marukawa et al. [9] have proposed two methods of combining character recognition for retrieving Japanese documents. In their recognition process they used a multi-template based on directional features. The segmented character pattern is normalized and the contour derived from this geometrically normalized binary pattern is represented by eight directional codes. Each directional code is mapped into one of the four feature patterns. Each of them corresponds to the horizontal, vertical, right-up, and left-up directions. These features are blurred into a "directional feature" pattern.

Their first method [9] is the Extended Query Term Method using (Method I) confusion matrix, which uses two steps. In the first step characters similar to a character in the query terms by using a confusion matrix, and strings combining the "similar characters" are expanded as new query terms. In the next step new query terms are created by supplemented similar characters. In their second method, (Method II) [9] non-deterministic text, which keeps multiple candidates in a text file, is used. Searching is done by using clean query terms

5 Bi-gram and Its Application to Online Character Recognition

Because Japanese language has a huge character set including characters with different entropies it is difficult to apply conventional methodologies based on n-gram to post-processing in Japanese character recognition. Itoh [10] proposed a method to overcome these two problems using a clustering scheme based on different parts of speech of Kanji and also by homogenizing the entropies of different Kana and Kanji characters. A bi-gram approach was used, based on these two techniques to Japanese language model. Experiments resolved the imbalance between Kana and Kanji characters, and reduced the perplexity of Japanese to less than 100, when Japanese newspaper texts were used. A post-processing technique was proposed using the model for on-line character recognition and about half of all substitution errors were obtained when the correct characters were among the candidates.

This approach needs to be extended to Katakana characters. Among the different parts of speech, verbs and post positions were considered extensively, but more minute classification of nouns were required. Finally the language model was applied as an online optical character recognition post processing method, caused failures in cases where the correct character is not included in the candidates. Integration of the language model into the recognition methodology should be attempted.

6 Offline Character Recognition Using Virtual Example Synthesis

In character recognition, both for printed and handwritten character recognition, the performance of classifiers strongly depend on quality of naming samples. A very large database containing a sufficiently large number of good examples are required for classifiers to perform well, particularly in the case of hand written character samples. This is costly and time consuming. Miyao and Maruyama [11] attempt to overcome this difficulty by synthesizing virtual examples from a small number of real samples. Their approach is implemented in two steps.

Their results indicated that with an appropriate number of eigenvectors and base samples, the recognition rates are higher than or equal to those without PCA based pattern segmentation and the classification time is faster as the support vector of support vector machine is further reduced for recognition of handwritten Hiragana characters due to (I) determination of cumulative recognition rate for improvement in effectiveness and (II) designing a decision directed acyclic graph based on support vector machines (SVM).

7 Inter-word Spacing in Japanese

As mentioned in section 2, Japanese words are not separated by delimiters like spaces, thus making character recognition difficult. Although Japanese is a word based language, segmenting text into word is not as clear cut as in languages using word spacing as a rule. Spacing is incorporated as in at least two ways. The first way is by adding spaces not only between their grammatical modifiers and post positions. The second way is to consider the modifiers and post positions as a part of the modified word. Based on the study conducted by Saino et al. [12] using 16 subjects in Japanese reading, 60 word texts from excepts of newspapers and internet columns, it was concluded that in pure Hiragana text, inter-word spacing is an effective segmentation method, in contrast to Kanji-Hiragana text, since visually silent kanji characters serve as effective segmentation uses by themselves.

8 Character Features Vectors Identification

Every character has its own features and identities. By identifying features we can recognize characters from a textual image document. By feature extraction the critical characteristics of characters gets isolated, and that reduces the complexities of the pattern. After classification it compares with known patterns and then matched with the character that has the same characteristics.

Barners and Manic [16] proposed an algorithm that contributes an original approach to constructing feature vectors; their proposed methodology creates a neural network by design and not by training. They showed that the center of gravity remains consistent even a character is rotated. The center of gravity will move proportionally as the characters change in size, translation, or rotation. The dakuon and handakuon characters will also be accurately identified due to the presence of the dakuten and handakuten markers. The center of gravity moves proportionally with the additional pixels and produces a set of unique feature characteristics.

The Size-Translation-Rotation-Invariant Character Recognition and Feature vector Based STRICR-FB algorithm is based on the Kohonen Winner Take All [13, 14], type of unsupervised learning. The algorithm comprises of two phases; Construction of Character Unique Feature Vectors which calculates distance between characters and in an expanded form of the Euclidean distance defined in (15). The next phase is passing character unique feature vectors through a neural network for character recognition.

They conducted three list sets to validate the algorithm. In general, a training set is used to create an artificial neural network (artificial neural network). A test set of random characters is then used to determine the effectiveness of this artificial neural network. The experiment by random characters produces three sets of results; among which rotation set produced 96.2% accuracy rate and that of random character set is 93%.

9 Performance Improvement Strategies on Template Matching

Handwritten text differs due to differences in writing styles, and hence, handwritten character recognition suffers from absorbing variations of the same characters among

different writing styles. Also, resolution of the graphical similarity of different characters in Japanese text is another consideration to be taken into account. To overcome the problem, an offline effective algorithm for large scale character recognition for large set characters like Korean and Chinese was proposed by Kim [17]. The algorithm was developed based on template matching and improvement strategies; First, Multi-stage pre-classification that reduces the processing time of the template matching by cutting off a number of recognition target classes [18] is done. It is desirable to cut off as many classes as possible with little or no degradation of recognition accuracy. Second, the pair wise reordering is done to enhance the recognition accuracy by performing a fine detail classification on the recognition candidates generated from the template matching [19].

The resulting algorithm consists of three processing stages of multi-stage pre-classification, template matching and pair wise reordering. The algorithm showed its effectiveness by an experiment where handwritten Korean character came up with 86.0% of recognition accuracy and 15 characters per second from PE92 [20] handwritten Korean character database.

10 Proposed Recognition Procedure

A scanned copy of a hand written text was taken and followed by 3 image processing steps, namely, 1) segmentation, 2) recognition and 3) cluster in to groups. Later the recognition process was followed up by character resizing, extracting features, group identification and learning recognition. The resulting image might contain noise and hence a Gabour filter had been used. Later we had identified the group and devised a learning recognition system using neural network. The process flow chart is as following:

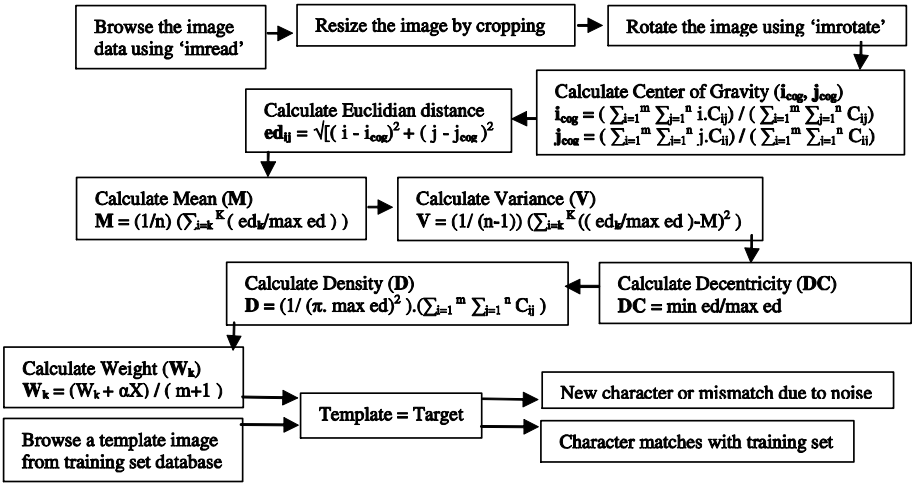


Fig. 5. The flowchart has been made considering the Japanese characters could be identified.

11 Results

We have stepped down the out comes into three parts, 1) Segmenting the image for preprocessing, 2) identifying the features following STRICR-FB algorithm, 3) template and target image matching. The following are the interfaces of the application of step-1 and step3, used in MATLAB simulator.

11.1 Image Rotation Resizing and Making Ready for Feature Extraction

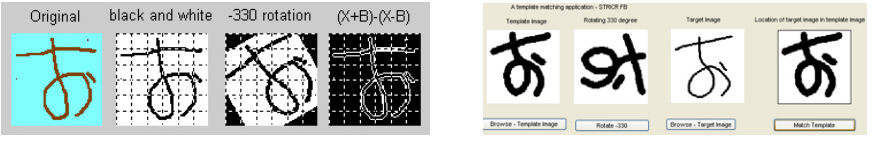


Fig. 6. Interface (Fig. left) for Image binary conversion, resizing, rotation & COG calculation. And the figure (Fig. right) shows template and target image matching interface.

11.2 Identifying the Features and Clustering and Target-Template Interface

Finding COG (Center of Gravity), Mean, Variance, Density and Decentricity and Euclidean distance in a function in MATLAB for each of target pictorial Japanese character have been done in this phase. This section is under progress, the interface has been show in future work section. The Figure (Fig 6. right) shows the result of interface for matching target and template images, accurately getting matched.

12 Result Analysis and Future Work

The following results of mismatched character after rotation and the future interface of corrected application have been shown together.

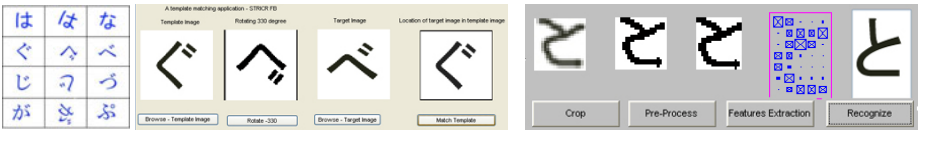


Fig. 8. The table (left) shows the Japanese characters after rotation and getting matched with some other Japanese characters, and the interface of the application fails to match the character. The figure (middle) shows the wrongly matched between rotated template character and the target one. We took gu(ぐ) and be(べ) as template and target images receptively. And the figure (right), shows the final and future window for recognizing Japanese hiragana “to” (と).

Even if we are still trying to identify the characters, but after rotating many of those Japanese characters, identification was not done properly. This was due to similar shaped Japanese characters and the risk of getting matched with rotating shape and another Japanese character. Also the reason behind this failure is also considered due

to the presence of the noise. Reduction of noise using Gabour filter will reduce down tendency of getting wrongly matched.

We are trying to figuring out (Fig.8. right) some new more features related to characters of the Japanese text and its architectures. The interface of the Japanese character identification is what under progress. Since there was no computation of time and space complexities in earlier methodologies but we can add a timer in future.

13 Comparison with Earlier Methodologies

We have been working on first five Japanese Hiragana characters only, except dakuon and handakuon strokes. And the result worked well, it showed almost 90% of accuracy. There are four characters; we had worked on, for template target dissimilarity.

Table. 9. The table above shows the comparison among the methodologies for identifying Japanese characters uniquely

Methods	Approaches	Error	Correction
Error tolerant Search System	OCR for multiple ch. set, Transmedia Machine, Segmentation ambiguities.	Recognition error in document processing system.	Enhance performance, remove noise.
Document Retrieval Tolerating Character Recognition Errors	Extended Query Term, non-deterministic text.	Complexities in measuring Extended Query Term, Couldn't conduct retrieval experiments with complex query.	Not compared with complex query condition, Not relevance in ranking capability.
Bi-gram - Online	Clustering ch. set a/c to the parts of speech. Classifying it into more detailed sub-categories with part-of-speech attributes.	Difficulty in applying conventional methodologies, post processing.	Extension for katakana is required.
Virtual Example Synthesis - Offline	Synthesizing virtual attempts (from small, real sample),	Cost and time due to large and proper database	Enhance performance, remove recognition rate.
STRICR-FB	Identifying features COG, Euclidean dist. Mean, Variance, density, Decentricity.	Problem in identifying characters due to size and rotational changes.	Noise, uniquely identified features for distinguishing i/p and template image.

14 Concluding Remarks

Document Retrieval in Japanese text is complicated by three main features. Firstly, the text comprises of over 3000 characters based on syllabic characters (Kana) and ideographic character (Kanji). Secondly, there are no spaces between two words in Japanese text. This can create problems in pure Hiragana text, though it is not so much problem in Kanji – Hiragana text due to the presence of visually salient Kanji characters. Thirdly, homomorphism or similar shape definition for several Japanese characters could yield errors in the recognition some of these issues have been addressed in the articles included in the survey.

It is hoped that with use of proper IP and character vector techniques, the document retrieval process in Japanese text can be conducted, if net eliminated.

References

1. Dahl, D.A., Norton, L.M., Taylor, S.L.: Improving OCR accuracy with linguistic knowledge. In: Proc. Second Ann. Symp. Document Analysis and Information Retrieval, pp. 169–177 (1993)
2. Niwa, N., Kayashima, K., Shimeki, Y.: Postprocessing for character recognition using keyword information. In: IAPR Workshop Machine Vision Applications, pp. 519–522 (1992)
3. Hull, J.J., Li, Y.: Word recognition result interpretation using the vector space model for information retrieval. In: Proc. Second Ann. Syrup. Document Analysis and Information Retrieval, pp. 147–155 (1993)
4. Tanaka, Y., Torii, H.: Transmedia machine and its keyword search over image texts. In: Proc. RlAO 1988, pp. 248–258 (1988)
5. Trenkle, J.M., Vogt, R.C.: Word recognition forinformation retrieval in the image domain. In: Proc. Second Ann. Symp. Document Analysis and Information Retrieval, pp. 105–122 (1993)
6. Hull, J.J.: Document image matching and retrieval with multiple distortion–invariant descriptors. In: Proc. IAPR Workshop on Document Analysis Systems, pp. 383–399 (1994)
7. Fujisawa, H., Hatakeyama, A., Nakano, Y., Higashino, J., Hananoi, T.: Document storage and retrieval system. U.S. Patent 4985863 (1986)
8. Senda, S., Minoh, M., lkeda, K.: Document image retrieval system using character candidates generated by character recognition process. In: Proc. Second Int. Conf. Document Analysis and Recognition, pp. 541–546 (1993)
9. Marukawa, K., Hu, T., Fujisawa, H., Shima, Y.: Document retrieval tolerating character recognition errors—evaluation and application. Pattern Recognition 30(8), 1361–1371 (1997); Oriental Character Recognition
10. Itoh, N.: Japanese language model based on bigrams and its application to on-line character recognition. PR 28(2), 135–141 (1995)
11. Maruyama, K.-I., Maruyama, M., Miyao, H., Nakano, Y.: Handprinted Hiragana recognition using upport vector machines. In: Proceedings of Eighth International Workshop on Frontiers in Handwriting Recognition, pp. 55–60 (2002), doi:10.1109/IWFHR.2002.1030884
12. Sainio, M., Bingushi, K., Bertram, R.: The role of interword spacing in reading Japanese: An eye movement study. Vision Research 47(20), 2577–2586 (2007)
13. Kohonen, T.: Self-organized formation of topologically correct feature maps. Biological Cybernetics 43, 59–69 (1982)
14. Kohonen, T.: Self-Organization and Associative Memory, 2nd edn. Springer, New York (1988)
15. Hung, D., Cheng, H., Sengkhamyong, S.: Design of a Hardware Accelerator for Real-Time Moment Computation: A Wavefront Array Approach. IEEE Transactions on Industrial Electronics 46(1) (February 1999)
16. Barnes, D., Manic, M.: STRICR-FB, a Novel Size-Translation- Rotation-Invariant Character Recognition Method. In: 2010 3rd Conference on Human System Interactions (HSI), pp. 163–168 (May 2010)

17. Kim, S.H.: Performance Improvement Strategies on Template Matching for Large Set Character Recognition. In: Proc. 17th International Conference on Computer Processing of Oriental Languages, Hong Kong, pp. 250–253 (April 1997)
18. Tung, C.H., Lee, H.J., Tsai, J.Y.: Multi-stage pre-candidate selection in handwritten Chinese character recognition systems. Pattern Recognition 27(8), 1093–1102 (1994)
19. Takahashi, H., Griffin, T.D.: Recognition enhancement by linear tournament verification. In: Proc. 2nd ICDAR, Tsukuba, Japan, pp. 585–588 (1993)
20. Kim, D.H., Hwang, Y.S., Park, S.T., Kim, E.J., Paek, S.H., Bang, S.Y.: Handwritten Korean character image database PE92. In: Proc. 2nd ICDAR, Tsukuba, Japan, pp. 470–473 (1993)

Fast Fingerprint Image Alignment

Jaspreet Kour[1], M. Hanmandlu[2], and A.Q. Ansari[3]

[1] GCET, Knowledge Park Greater Noida, U.P. India
tojaspreet@gmail.com
[2] IIT, Hauz Khas, New Delhi, India
mhmandlu@gmail.com
[3] Jamia Milia Islamia, New Delhi, India
aqansari@gmail.com

Abstract. Fingerprint is one of the various modalities used in biometrics for authentication. An important issue, when designing a fingerprint–based biometric system / application is alignment of fingerprint images before feature extraction and matching. In this paper we present fingerprint alignment algorithm based on Principal Component Analysis (PCA). PCA based method is compared in terms of average time taken for fingerprint image alignment with the existing methods for fingerprint alignment. Experiments show that PCA based method is able to achieve alignment of fingerprint images in FVC2002 DB1A accurately and the algorithm is robust and fast.

Keywords: Biometrics, Fingerprint Alignment, PCA.

1 Introduction

Biometrics refers to identifying a person based on physiological or behavioral characteristics and has the capability to reliably distinguish between a genuine person and an imposter. Biometrics is reliable and more capable than the traditional knowledge based and token based techniques. Fingerprints are considered to be one of the most popular biometric authentication and verification measures because of their high acceptability and uniqueness. A fingerprint is the reproduction of a fingertip epidermis, produced when a finger is pressed against a smooth surface. The most evident structural characteristic of a fingerprint is a pattern of interleaved ridges and valleys, in a fingerprint image [1].

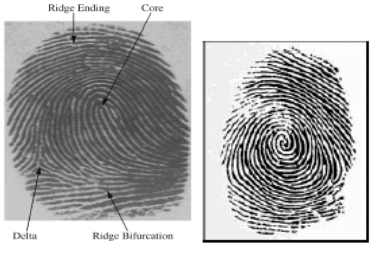


Fig. 1. Fingerprint Image

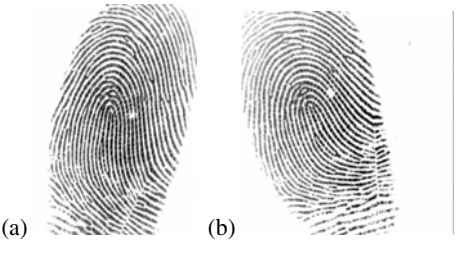


Fig. 2. Fingerprint image (a) rotated clockwise and (b) rotated anticlockwise

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 93–99.
springerlink.com

For comparison of two fingerprints, a feature vector consisting of discriminating features is extracted from the fingerprint image and matched against each other. Reliably matching fingerprint images is an extremely difficult problem, mainly due to the large variability in different impressions of the same finger (i.e, large intra-class variations)[1]. The performance of automated fingerprint recognition system is greatly affected by fingerprint translation and rotation. Fingerprint translation problem can be solved by extracting region of interest around a reference point called core point. Fingerprint rotation refers to the alignment of a captured fingerprint image when measured against the image background [2]. Rotation of fingerprint image can be clockwise or anticlockwise relative to the vertical axis of the fingerprint image as shown in Figure 2.

In the automated fingerprint matching systems, an efficient and accurate alignment algorithm plays a crucial role in the performance of the system. Different alignment algorithms are based on different features of fingerprint images. When analyzed at the global level, fingerprints exhibit macro details such as ridge flow and pattern types, orientation field, ridge frequency etc and the important features at the local level, called minutiae can be found in the fingerprint patterns. Minutiae means small detail; in the context of fingerprints it refers bifurcations and ridge endings.

The organization of this paper is as follows: Section 2 gives a review of some fingerprint alignment methods. Basic concepts of Principal Component Analysis are discussed in Section 3. Section 4 describes fingerprint alignment based on PCA. Section 5 discusses database used and experimental results obtained. Conclusion and future work are presented in Section 6.

2 Literature Review

Realignment of a fingerprint involves the use of some landmark as a reference point. The realignment angle and direction relative to vertical axis can be computed after finding the chosen reference point. The True Fingerprint Centre Point (TFCP) is the chosen reference point and is located in the fingerprint mask. The realignment direction (clockwise or anticlockwise) is found from mask. Realignment angle is obtained and fingerprint image is rotated upright [2]. The orientation, scale and translation parameters are estimated using a generalized Hough transform for fingerprint registration in [3]. Fingerprint registration and alignment based on minutiae are widely used. Alignment is an important step as minutiae extraction algorithms and minutiae matching algorithms are dependent on the alignment of fingerprint images. These methods are accurate but computationally expensive. Owing to spurious minutiae extracted from low quality images, such methods may result in false alignment and matching. The local and global structures of fingerprint are used in [4]. Local structures are used directly for matching and the best matched local structures provide the correspondence for aligning the global structure of the minutiae. Use of Genetic Algorithm is made for optimizing the alignment of a pair of fingerprint images. GA based registration is reported to be ten times faster than 3D algorithm with similar alignment accuracy and 13% more accurate than 2D algorithm with the same running time. Alignment of two fingerprints position is obtained by certain landmarks such as core points and the translation and rotation parameters are obtained by comparing the

coordinates and orientation of two core points [6]. A number of fingerprint kernels for different classes of fingerprints is defined in [7] and the best fit kernel to classify fingerprint image is found. Location and orientation of two kernels are used to align the query and enrolled fingerprint image. A RANSAC based method to determine a rigid transformation which aligns two fingerprint images using solely minutiae coordinates and angles is presented in [8]. A ring model to align a pair of fingerprint images based on single singular points is proposed in [9].The alignment algorithm is using only singular point and the direction information around it to translate and rotate the input fingerprint image. The region far from the singular point is not well aligned due to nonlinear distortion. In [10] a two stage optimization alignment which combines the global and local features is discussed. The initial registration is done using global features such as orientation field, curvature maps and ridge frequency maps. In the next step local features such as minutiae are used for fine tuning of transformation parameters obtained from the initial step. A pre alignment algorithm is discussed in [11] and it does not require huge memory for storage in the smart card and the information sent out is not confidential. Five regions from the fingerprint are extracted and these regions are coded as a triplet. Pre-alignment starts by rotating five regions and correlating them with the resulting image. Eight types of special ridges are introduced to align two fingerprints in [12]. The ridge with the maximum curvature is used as a reference ridge for the initial alignment and the corresponding special ridges paired by topology get aligned. Alignment using the special ridges is fast and robust. In [13] k-means and Fuzzy c-means have been used for aligning fingerprint images for rotation invariance.

3 Basics of Principal Component Analysis (PCA)

Principal component analysis (PCA) is one of the statistical techniques frequently used in signal processing for the dimension reduction [14]. Principal component analysis (Karhunen-Loeve or Hotelling transform)-PCA belongs to linear transforms based on the statistical techniques. This method provides a powerful tool for data analysis in signal and image processing [15]. Principal component analysis coverts an n-dimensional vector $x = [x_1, x_2, ...x_n]^T$ into a vector y according to

$$y=A(x-m_x) \tag{1}$$

The vector m_x in Eq. (1) is the vector of mean values of all input variables given by

$$m_x = E\{x\} = \frac{1}{K}\sum_{k=1}^{K} x_k \tag{2}$$

Matrix A in Eq. (1) is determined by the covariance matrix C_x. Rows in the A matrix are formed from the eigenvectors e of C_x ordered according to corresponding eigen values in descending order. C_x matrix is given by the relation

$$C_x = E\{(x-m_x)(x-mx)^T\} = \frac{1}{K}\sum_{k=1}^{K}(x_k x_k^T - m_x x_x^T) \tag{3}$$

As the vector x of input variables is n-dimensional it is obvious that the size of C_x is n x n. The elements C_x (i, i) lying in its main diagonal are the variances of x

$$C_x(i,i) = E\{(x_i - m_i)^2\} \tag{4}$$

The rows of A in Eq. (1) are orthonormal so the inversion of PCA is possible according to the relation

$$x = A^T y + m_x \tag{5}$$

The kernel of PCA defined by Eq. (1) has some properties from the matrix theory which can be used in the signal and image processing to fulfill various goals such as determination of object rotation [14].

Properties of PCA can be used for determination of selected object orientation or its rotation. Various method of image segmentation to object definition (like thresholding, edge detection or others) must be used at first. Binary image containing object boundary or its area in black (or white) pixels on the inverse background results from this process. After that two vectors a and b containing the cartesian x and y coordinates of object's pixels can be simply formed. The vector x in the Eq. (1) is in this case a 2-dimensional vector consisting of a and b respectively. The mean vector m_x and the covariance matrix C_x are computed as well as its eigenvector e. Its two elements - vectors e1 and e2 enable the evaluation of object rotation in the cartesian axis or object rotation around the center given by m_x [14].

4 PCA Based Fingerprint Alignment

Alignment is a crucial step in fingerprint matching algorithms as misalignment of the two fingerprints of the same finger certainly produce a false matching result. Owing to rotation, scaling and translation between the enrolled fingerprint in the database, alignment is a necessary step in fingerprint matching algorithms. PCA is mainly helpful in finding the directions along which spread of the data is more. The amount of spread is given by eigen values and the direction of spread is given by eigen vectors. Following steps are followed for alignment of fingerprint images.

1. Binarize the image and find the coordinates of the points at which fingerprint is located.
2. Find the mean of all these points and then find covariance matrix.
3. Find the eigen value and eigen vectors based on the covariance matrix.
4. Find the orientation angle based on eigen vector corresponding to maximum eigen value. This gives the angle of rotation.
5. If this angle of rotation is more than 60 degrees (as it is assumed that fingerprint will not have rotation more than 60 degrees), then find the angle of rotation based on other eigen vector.
6. Rotate the fingerprint by rotation angle obtained.

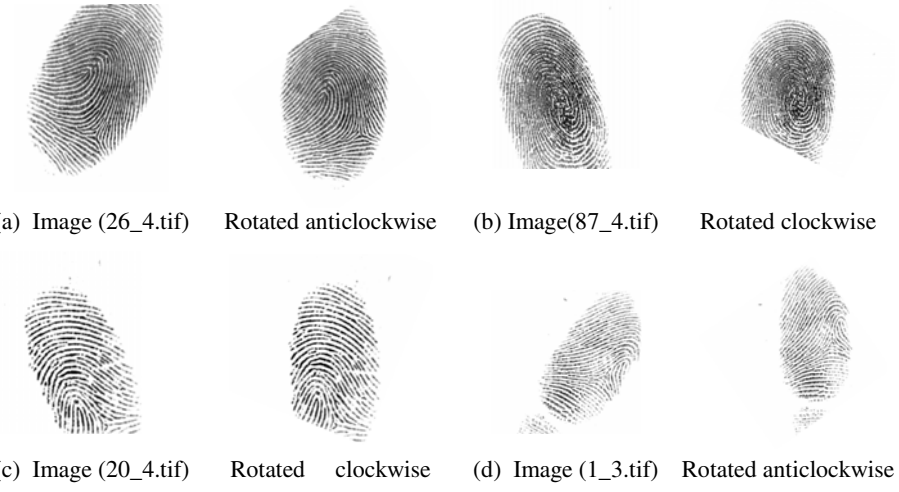
5 Experimental Results

The rotation of fingerprints, during the enrolment acquisition by the scanner, can lead to false rejection phenomenon where a legitimate subject is classified as an imposter [2].

The experiments reported in this paper have been conducted on the benchmark fingerprint database DB1A in FVC2002 [16]. It comprises 800 fingerprint images of size 388 x 374 pixels captured by optical sensor at a resolution of 500dpi, from 100 fingers (eight impressions per finger).

5.1 Alignment Results Based on PCA

Third and fourth sample of all the individuals in the database have rotation either in the clockwise or the anticlockwise direction[16]. PCA based alignment algorithm is applied on the rotated images of FVC2002 DB1A and checked for rotation correction. The results of alignment are shown in Fig. 2.



(a) Image (26_4.tif) Rotated anticlockwise (b) Image(87_4.tif) Rotated clockwise

(c) Image (20_4.tif) Rotated clockwise (d) Image (1_3.tif) Rotated anticlockwise

Fig. 2. Original images and rotated images obtained after alignment

Time taken for the alignment of five images from the database is shown in Table 1. PCA based method takes an average time of 0.556 sec. Time taken by each image using PCA based algorithms is shown in Fig 3.

Table 1. Average time taken for PCA Alignment

Fingerprint images from FVC2002 DB1A	Time taken for alignment (sec.)
1_3.tif	0.492
13_4.tif	0.573
39_4.tif	0.573
87_4.tif	0.569
93_4.tif	0.571
Average time	**0.556**

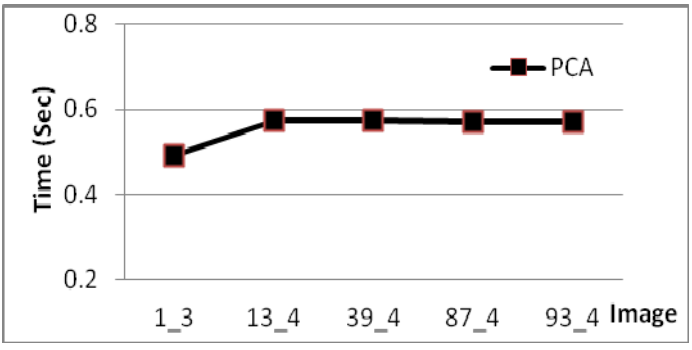


Fig. 3. Run time of PCA based alignment

5.2 Comparison with the Existing Methods

The PCA based method for fingerprint alignment is compared with the existing methods in terms of average time taken for aligning the images. The average time taken for alignment by registration based method is 17.6 sec, 0.946 sec by k-means algorithm and 2.558 sec by Fuzzy C-means algorithm [13]. Comparison of the proposed alignment algorithm with the existing algorithms in terms of average time taken for alignment is shown in Fig 4.

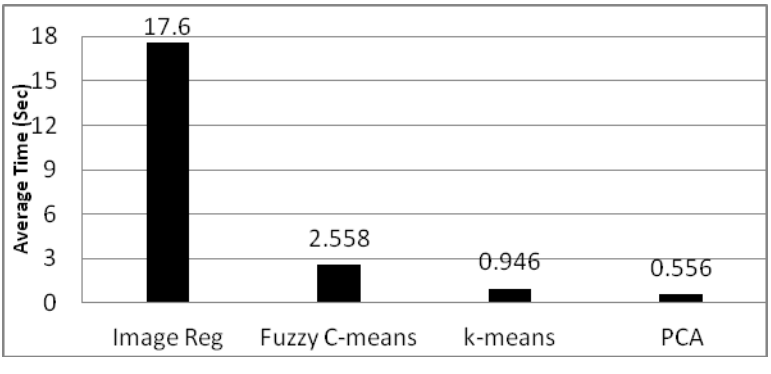


Fig. 4. Comparison with existing methods

6 Conclusion

A fingerprint alignment algorithm based on PCA is presented. From the experimental results it has been observed that PCA based alignment approach is able to align the rotated images in the database accurately and also the average time taken by it for aligning fingerprint images is less. PCA based alignment algorithm is compared with the existing algorithms for fingerprint alignment. PCA based alignment method is robust and fast. The improvement in recognition accuracy with and without proposed fingerprint alignment algorithms is considered as the future work.

References

1. Maltoni, D., Maio, D., Jain, A.K.: Handbook of Fingerprint Recognition. Springer, New York (2003)
2. Ishmael, S., Tendani, M., Malumedzha, C., Leke-Betechuoh, B.: A Novel Fingerprint Realignment Solution that Uses the TFCP as a Reference. International Journal of Machine Learning and Computing (IJMLC) 01(03), 297–304 (2011)
3. Ratha, N.K.: A real–time matching system for large fingerprint databases. IEEE Transactions on Pattern Analysis and Machine Intelligence 18, 799–813 (1996)
4. Jiang, X., Yau, W.Y.: Fingerprint Minutiae Matching Based on the Local and Global Structures. In: Proceedings International Conference on Pattern Recognition, vol. 2, pp. 1042–1045 (2000)
5. Ammar, H.H., Tao, Y.Y.: Fingerprint registration using genetic algorithm. In: Proceedings of 3rd IEEE Symposium on Application Specific Systems and Software Engineering Technology (2000)
6. Nilsson, K., Bigun, J.: Prominent symmetry points as landmarks in fingerprint images for alignment. In: Proceedings 16th International Conference on Pattern Recognition, pp. 395–398 (2002)
7. Jain, A.K., Minut, S.: Hierarchical kernel fitting for fingerprint classification and alignment. In: Proceedings International Conference on Pattern Recognition, pp. 469–473 (2002)
8. Ramoser, H., Wachmann, B., Bischof, H.: Efficient alignment of fingerprint images. In: Proceedings International Conference on Pattern Recognition, pp. 748–751 (2002)
9. Li, F., Maylor, K.H., Leung, Liu, C.: Fingerprint Alignment Using Ring Model. In: Proceedings of the Third International Conference on Information Technology and Applications, Sydney, Australia, vol. 1, pp. 738–743 (2005)
10. Yager, N., Amin, A.: Fingerprint alignment using a two stage optimization. Pattern Recognition Letters 27, 317–324 (2006)
11. Lam, H.K., Yau, W.Y., Chen, T.P., Hou, Z., Wang, H.L.: Fingerprint pre-alignment for hybrid match-on-card system. In: 6th International Conference on Information, Communications & Signal Processing, Singapore, pp. 1–4 (2007)
12. Hu, C., Yin, J., Zhu, E., Chen, H., Li, Y.: Fingerprint alignment using special ridges. In: 19th International Conference on Pattern recognition, ICPR, Tampa, FL, pp. 1–4 (2008)
13. Jaganthan, P., Rajinikannan, M.: Fast Fingerprint Image Alignment Algorithms Using K-Means and Fuzzy c-Means clustering based image rotation. In: Proceedinigs of First International Conference on Logic, Information, Control and Computation, ICLICC, Gandhigram, India, pp. 28–288 (2011)
14. Mudrová, M., Procházka, A.: Principal component analysis in image processing
15. Gonzales, R.C., Woods, R.E.: Digital Image Processing, 2nd edn., p. 795. Prentice Hall (2002) ISBN 0-201-18075-8
16. FVC (2002), `http://bias.csr.unibo.it/`

Colour and Texture Feature Based Hybrid Approach for Image Retrieval

Dipti Jadhav, Gargi Phadke, and Satish Devane

Ramrao Adik Institute of Technology, Navi - Mumbai. India
{dipti.jadhav,gargi.phadake,satish}@rait.ac.in

Abstract. The Content Based Image Retrieval (CBIR) is a technique that works on images and in response extracts relevant images. A novel hybrid two stage universal CBIR technique using both colour and texture features extraction is proposed in this paper. In the first stage for colour feature extraction, colour moments up to the fourth order are extracted and are used in deriving the respective histograms which forms the colour feature vector. In the second stage for the texture feature extraction the CCM (Colour Co-occurrence Matrix) technique employed takes into account the correlation between the RGB colour bands in all the eight directions while computing the texture features. In every stage the distance between the query image and the image in the database is calculated by using relative distance measure. The resultant distance between the query image and the image in the database is calculated by using a weighted distance classifier. Thus, a hybrid fusion method is achieved that has better performance than other colour-spatial based methods and promises to give more relevant output to the user.

Keywords: CBIR, local statistics histograms, Skew, Kurtosis, CCM.

1 Introduction

The recent tremendous growth in computer technology has brought a substantial increase in the storage of digital imagery. Examples of applications can be found in everyday life, from museums for archiving images or manuscripts, to medicine where millions of images are generated by radiologists every year. Storage of such image data is relatively straightforward, but accessing and searching image databases is intrinsically harder than their textual counterparts. The solution to this is CBIR. The goal of CBIR systems is to operate on collections of images and, in response to visual queries, extract relevant images.

Image low-level visual features as well as high-level semantic features are being used by CBIR systems [1], [2]. Image content features such as colour, texture, shape, etc., which are analyzed and extracted automatically by computer achieves the effective retrieval [3], [4]. The cumulative colour histogram was proposed by M. Strick, M. Orego [4]. G Pass and R Zabih [5] proposed a technique for comparing images called histogram refinement, which imposes additional constraints on histogram based matching. J. Huang [6] proposed a feature called Colour Correlogram which is used to express how the spatial correlation of pairs of colours

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 101–111.
springerlink.com

changes with distance. A. Rao, R. Srihari, and Z. Zhang [7], presented a feature called Annular Colour Histogram (ACH), used to express the distribution of each identical colour bin in concentric circles centred at the centroid of the bin with different radiuses. L. Cinque, et al. [8] introduced Spatial-Chromatic Histogram (SCH) a new indexing methodology for integrating colour and spatial information for CBIR. S. Lim, G Lu [9] introduced Geographical Statistics (Geostat) method to describe the spatial distribution of identical colour with one parameter of "Looseness". Nidhi Singhai and S.K Shandilya [13], presented a survey on CBIR systems.

Most of the methods referred above have taken into consideration only the colour feature for image retrieval which is found to be sensitive to image rotation and translation, but have not taken into consideration the texture feature, which has a profound impact on image retrieval. Most of the images are a good composition of colour and texture, hence we are proposing a two stage hybrid CBIR system which considers the influence of the colour factor as well as the texture factor and the results has been found quite satisfactory.

2 Proposed Methodology

The proposed system works in two stages to collect the quantitative information about the behaviour of the pixels. The first stage measures the behaviour of the distribution of the colours in the image. The second stage measures the co-relation of the pixels with its neighbors. Thus the information collected helps in enhanced retrieval.

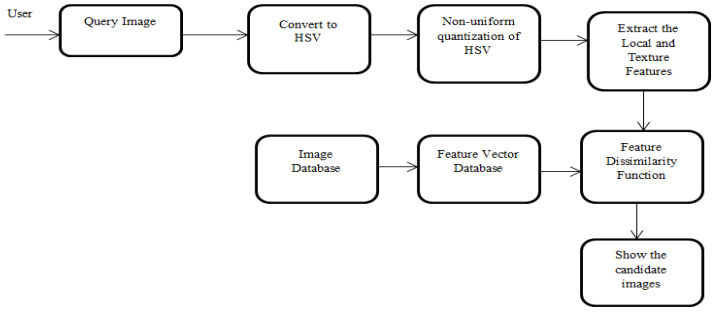


Fig. 1. Block diagram of the proposed system

For the proposed system the input image is a RGB image which is non-uniformly quantized into 166 bins after its conversion to HSV space. The database is a collection of colour images having different characteristics. For each image in the database its colour and texture feature vectors are retrieved. The colour and texture features of the query image are calculated. Feature dissimilarity between the query image and the database images is computed. The system then outputs the most relevant images to the user in the descending order. The block diagram is given in Fig. 1 and the detailed explanation of every block is given in the following sections.

3 Colour Feature Extraction

The RGB query image submitted by the user is initially converted to HSV space. Non-uniform quantization is performed to quantize the image into 166 bins. The values obtained will be the h (hue),s(saturation),v(bright)component values, h ϵ [0, 2π], s ϵ [0, l], v ϵ [0, l] [10]. The colours with v<0.25 are coded as q=0. The colours with s<0.2 and v>=0.25 are uniformly classified into 3 gray- levels according to the values of v, they are coded as q=l, 2, 3, respectively. All remaining colours (s>=0.2 and v>=0.25) are quantized into 162 colour bins, the codes of them are: $q = 9 \times h + 3 \times s + v + 4$, where h ϵ [0, 2π] be uniformly quantized into 18 bins with h=0, 1, ..., 17. s ϵ [0.2, 1] be uniformly quantized into 3 bins with value. s=0, 1, 2. v ε [0.25,1] be uniformly quantized into 3 bins with value v=0,1,2. The quantized colour set is denoted by Q= {0, 1, 2….1 65}.

3.1 Local Colour Feature Histogram

The normalized quantized colour histogram is calculated in equation (1) for the colour query image f of size m by n pixels, for the quantized colour q at location (i, j), i.e. q =f (i, j)[10]

$$H_c(q) = \frac{1}{mn} \left(\sum_{i}^{m} \sum_{i}^{n} \delta \; (f(i, j) - q)) \; q \in Q \right) \tag{1}$$

where δ is the unitary impulse function.

3.2 Local Moments and Features Histogram

The local spatial statistic features of every pixel are calculated. Mean, variance, skew and kurtosis (first, second, third and fourth central moments) of the pixels in the 5X5 pixel neighbourhood is calculated as in equations (2) to (5) respectively.

The mean –

$$e(i, j) = \frac{1}{25} \sum_{k=i-2}^{i+2} \sum_{j-2}^{j+2} f(k, l) \tag{2}$$

The variance -

$$\sigma(i, j) = \left[\frac{1}{25} \sum_{k=i-2}^{i+2} \sum_{l=j-2}^{j+2} (f(k,l) - e(i, j))^2 \right]^{1/2} \tag{3}$$

The skew -

$$s(i, j) = \left[\frac{1}{25} \sum_{k=i-2}^{i+2} \sum_{l=j-2}^{j+2} (f(k,l) - e(i, j))^3 \right]^{1/3} \tag{4}$$

The kurtosis –

$$k(i,j) = \left[\frac{1}{25} \sum_{k=i-2}^{i+2} \sum_{l=j-2}^{j+2} (f(k,l) - e(i,j))^4 \right]^{1/4} \quad (5)$$

The above values are in range Q={0, 1, 2... 165}. Thus, calculating these distributions, we can get normalized local mean histogram, normalized local standard deviation histogram, normalized local skew histogram and normalized local kurtosis histogram as given in equations (6) to (9) respectively.

$$H_{ne}(e) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \delta \ (e(i,j) - e)\ e \in Q \quad (6)$$

$$H_{ne}(\sigma) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \delta \ (\sigma(i,j) - \sigma)\ \sigma \in Q \quad (7)$$

$$H_{ne}(s) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \delta \ (s(i,j) - s)\ s \in Q \quad (8)$$

$$H_{ne}(k) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \delta \ (k(i,j) - k)\ k \in Q \quad (9)$$

Above five histograms are the local colour feature vector of the image and is denoted as F_C.

4 Texture Feature Extraction

The calculation of the colour spatial statistic features of the image is followed by the texture feature extraction by using CCM (Colour Co-occurrence Matrix).

4.1 Texture Feature Extraction Based on CCM

CCM (Colour Co-occurrence Matrix) is an extension of Gray-level Co-occurrence Matrix (GLCM) which is used to find the texture features of gray-level images. CCM is commonly used for colour images. CCM expresses the texture feature according to the correlation of the couple pixels gray-level at different positions.

In texture feature extraction, the original RGB query image is separated into Red, Green, Blue channel. CCM matrices are developed for RR, GG, BB, RG, RB, GB with distance as one pixel and considering all the eight directions. The eight matrices for one pair of channel eg. RR is added to obtain a single matrix for that pair of channel. Four features are calculated from each CCM matrix. These features [12] are energy, contrast, entropy, and inverse difference as in equations (10) to (13) respectively:

$$\text{Energy } E = \sum_i \sum_j f(i,j)^2 \tag{10}$$

$$\text{Contrast I} = \sum_i \sum_j (i-j)^2 f(i,j) \tag{11}$$

$$\text{Entropy } S = -\sum_i \sum_j f(i,j) \log f(i,j) \tag{12}$$

$$\text{Inverse difference H} = \sum_i \sum_j \frac{1}{1+(i-j)^2} f(i,j) \tag{13}$$

Thus in this way a total of twenty-four texture features (6 matrices and from every matrix 4 values i.e. 6X4) are extracted which will form the texture feature vector for the image. Thus, we get the texture feature vector as the CCM texture features. i.e.

$$F_T = F_{CCM} \tag{14}$$

5 Distance Dissimilarity

5.1 Distance between the Colour Feature Vectors

The spatial colour similarity between the query image and the image in the database can be calculated by using the relative distance measure between the histograms or between the colour feature vectors of the query image M and the database image I as given in equation (15):

$$d1 = \sum_{q=Q} \frac{\left| H_C^M(q) - H_C^I(q) \right|}{1 + H_C^M(q) + H_C^I(q)} \tag{15}$$

Similarly we can define d2*(M,I),d3(M,I),d4(M,I) ,d5(M,I))* as the distances between two images mean, variance, skew, kurtosis histograms respectively. Finally the colour feature vector distance between the two images can be calculated as:

$$D_{colour} = F_C(Fc_M, F_{CI}) = ds(M,I) = \sum_{i=1}^{5} d_i(M,I) \tag{16}$$

5.2 Distance between the Texture Feature Vectors

The distance between the texture feature of the query image M and the image in the database I is calculated as in equation (17).

$$D_{Texture} = F_{TM} - F_{TI} \tag{17}$$

5.3 Weighted Distance Measure

The final distance between the query image M and the database image I is given by the weighted distance formula as (18):

$$D(M, I) = \min(w_1 D_{colour}(F_{CM}, F_{CI}), w_2 D_{Texture}(F_{TM}, F_{TI})) \quad (18)$$

where D_{colour} and $D_{Texture}$ is the absolute distance between the colour feature vectors and the texture feature vectors of the query image and the database image respectively. Suppose the distance between the local features returns *m* images and the distance between the texture features returns *n* images then the weight w_1 and w_2 will be calculated as:

$$w_1 = \frac{1}{\sum 1+2+....+m}. \quad (19)$$

Similarly,

$$w_2 = \frac{1}{\sum 1+2+....+n}. \quad (20)$$

The final distance between the query image and database image is the distance with minimum of w_1 or w_2. .The feature with minimum weight is dominant and the same images are outputted to the user. Thus, the proposed system helps in outputting the images with the features which are dominant in query image.

The indexing of the resultant images is done on the basis of the minimum measure of dissimilarity from the query image. i.e. the images with minimum distance from the query image will be displayed first and then the images with more distance from the query image.

6 Experimental Result and Performance Analysis

The simulation of the proposed work is done in MATLAB by creating a GUI. We have used JPEG images of size 480X480 from the Internet to form the database. The images are divided into classes as flowers, water, garden, cars, and beach to name a few. The visual analysis of only two categories viz. water and beach are shown. The performance analysis of the proposed system is done by calculating the precision and recall [11] as in equations (21) and (22).

Precision = (number of relevant images retrieved)/ (total number of images retrieved). (21)

Recall = (number of relevant image retrieved)/ (total number of relevant images in the database). (22)

The Fig. 2,3& 4 shows the results of the query image as water. Fig.2 shows results from the colour feature extraction method. Fig. 3 shows the results from the texture feature extraction method. Fig. 4 shows the results from the hybrid method.

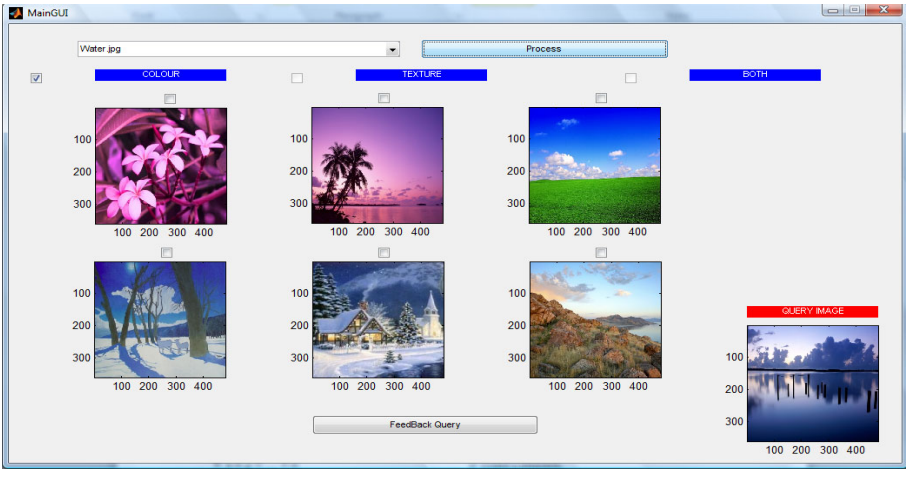


Fig. 2. Results for colour feature extraction method (Category Water)

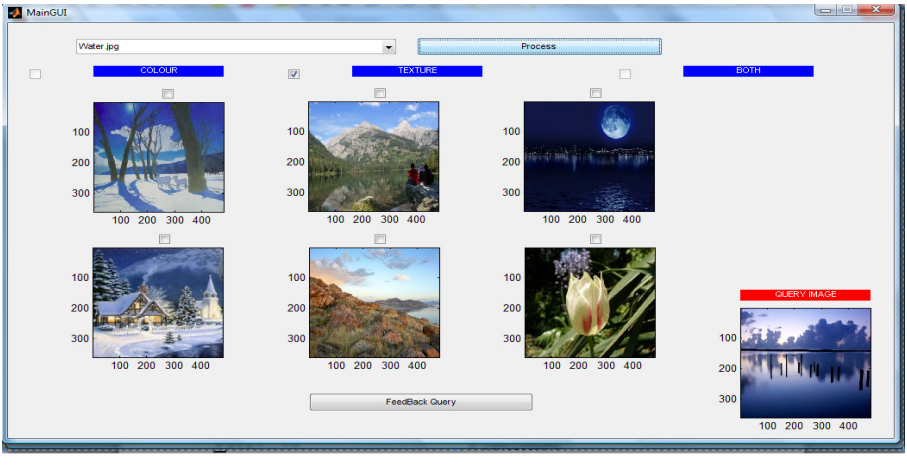


Fig. 3. Results for texture feature extraction method (Category Water)

Fig. 5, 6 & 7 shows the results of the query image as beach. Fig. 5 shows the results from the colour feature extraction method. Fig. 6 shows the results from texture feature extraction method. Fig. 7 shows the results from the hybrid method.

The proposed method helps to retrieve the images with the factor (colour or texture) that is more pronounced in the query image i.e if the query image contains the texture content dominantly, the retrieved images also contains images which have texture dominantly than colour content.

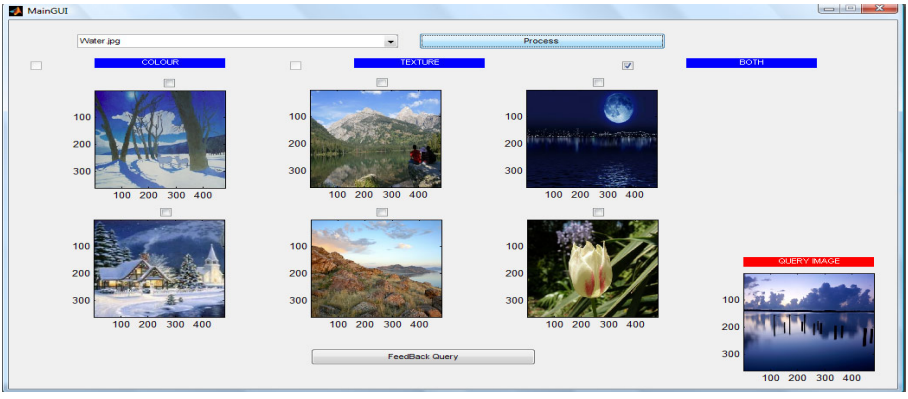


Fig. 4. Results for hybrid method (Category Water)

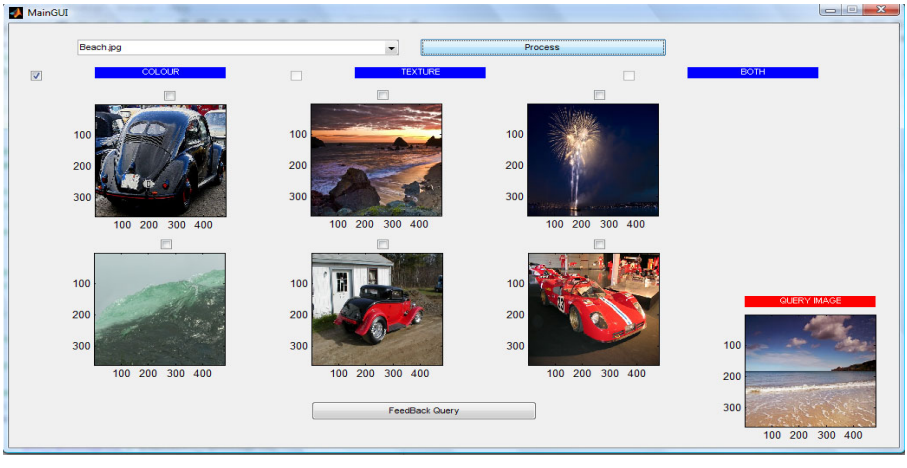


Fig. 5. Results of colour feature extraction method (Category Beach)

The database used here is a collection of random images from the Internet where the element of similarity between the images in the same category is very low. The backgrounds and the foregrounds of the images are quite different from each other. Suppose the category is water then the images in the database contains water as well as many other objects. Some images have very less water objects.

As can be seen in Fig. 4 and Fig. 7, the first 5 images have water but in small quantity as compared to the query image. Inspite of this the proposed system retrieves these images which have water objects. Thus this is the most appreciating application or advantage of the proposed system that it can be applied to any database which has a random set of images but still achieves good retrieval.

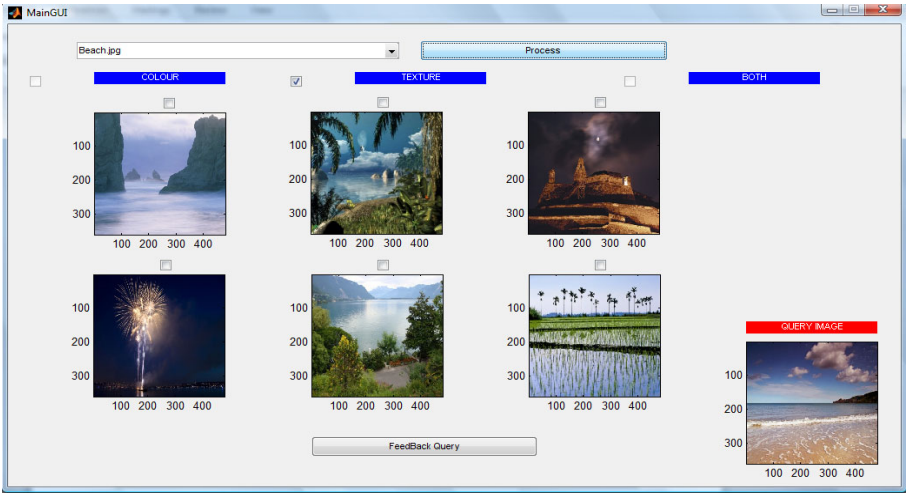


Fig. 6. Results of texture feature extraction method (Category Beach)

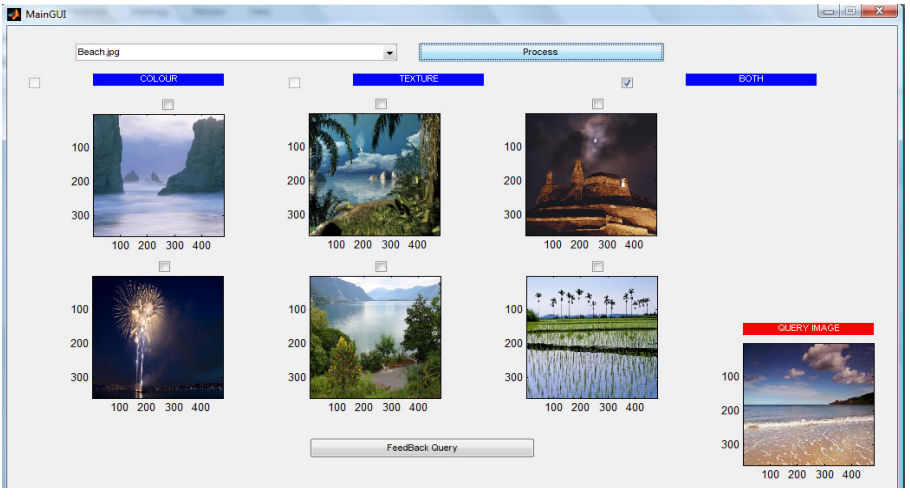


Fig. 7. Results of hybrid method (Category Beach)

6.1 Analysis

Table 1 gives the quantitative analysis for the proposed system for the two categories. It is seen that the proposed hybrid method gives better precision and recall than using a colour feature extraction or texture feature extraction method individually. The precision by using the texture feature extraction outperforms the colour method. The system gives appreciable precision and recall even when the amount of similarity between the images in the database is very low.

Table 1. Performance Analysis

Method	Category	Precesion	Recall
Colour Feature Extraction	Water	0.5	0.2
	Beach	0.5	0.2
Texture Feature Extraction	Water	0.67	0.3
	Beach	0.83	0.8
Hybrid Method	Water	0.7	0.35
	Beach	0.833	0.85

7 Conclusion

The proposed method works efficiently for different categories. In the colour feature extraction method since the third and fourth order moments are used, this method captures more colour spatial information in the image and makes the system irrelevant to image rotation and scaling. Since it is utilizing both colour and texture feature extraction it can capture more relevant information in the image. The future scope of the proposed work can be to reduce the semantic gap between the low-level and high-level semantic features. This can be done by appending a Relevance Feedback technique so that the output is much semantically closer to the users needs

References

1. Rui, Y., Huang, T., Chang, S.F.: Image retrieval: current technique, promising directions and open issues. Journal of Visual Communication and Image Representation 10, 39–62 (1999)
2. Wang, J., Wiederhold, G.: SIMPLTcity: semantics sensitive integrated matching for picture libraries. IEEE Trans. on Pattern Analysis and Machine Intelligence 23(8), 1–17 (2001)
3. Swain, M., Ballard, M.: Colour indexing. International Journal of Computer Vision 7, 11–32 (1991)
4. Strick, M., Orengo, M.: Similarity of colour images. In: Proc. of the SPIE2420. Storage and Retrieval for Image and Video Database III, San Jose, USA, pp. 381–392 (1995)
5. Pass, G., Zabih, R.: Histogram refinement for content-based image retrieval. In: Proc. of the 3rd IEEE Workshop on Applications of Computer Vision, Sarasota, pp. 96–102 (1996)
6. Huang: Image indexing using colour correlograms. In: Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Juan, pp. 762–168 (1997)
7. Rao, R.S., Zhang, Z.: Spatial colour histograms for content-based image retrieval. In: 11th IEEE International Conference on Tools With Artificial Intelligence (1999)
8. Cinque, S.L., Olsen, K., Pellicano, A.: Colour-based image retrieval using spatial-chromatic histograms. In: IEEE International Conference on Multimedia Computing and System, vol. 2, pp. 969–913 (1999)
9. Lim, S., Lu, G.: Spatial statistics for content based image retrieval. In: IEEE International Conference on Information Technology: Computers and Communications (2003)

10. Huang, C.B., Yu, S.-S., Zhou, J.-L., Lu, H.-W.: Image Retrieval Using Both Color and Local Spatial Feature Histograms. IEEE (2004)
11. Kong, F.-H.: Image Retrieval using both Color and Texture features. In: Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, July 12-15 (2009)
12. Arvis, V., Debain, C., Berducat, M., Benassi, A.: Generalization of Co-occurrence Matrix for Colour Images: Application to colour texture classification. Image Anal. Stereo 23, 63–72 (2004)
13. Singhai, N., Shandilya, S.K.: A Survey On: Content Based Image Retrieval Systems. International Journal of Computer Applications (0975 - 8887) 4(2) (July 2010)

Application of Software Defined Radio for Noise Reduction Using Empirical Mode Decomposition

Sapan H. Mankad[1] and S.N. Pradhan[2]

[1] Institute of Technology, Nirma University, Ahmedabad
sapanmankad@nirmauni.ac.in
[2] Institute of Technology, Nirma University, Ahmedabad
snpradhan@nirmauni.ac.in

Abstract. Software Defined Radios (SDR) are aimed at reducing the efforts required specifically in Wireless Communication. Many hardware devices are currently being used for communicating via radio waves. The function of SDR is to replace possibly all the hardware stuff by software which results in great flexibility and portability. This concept has opened new windows to the world of digital communication. Today there exists many flavors of SDR. This paper focuses on the open source GNU Radio and its capabilities. The GNU Radio Project serves as a reference for experiments in the area of signal processing and communications. This paper deals with utilizing the capabilities of software radios to improve the quality of the incoming signal. Our objective was to improve the received signal by reducing noise and thus enhancing the overall communication quality. We propose the use of Empirical Mode Decomposition (EMD) method embedded into GNU Radio. The idea presented here is to include the EMD functionality in GNU Radio toolkit so as to ensure reduction of error for better communication. We have integrated the capabilities of Empirical Mode Decomposition into GNU Radio and found improvements in the simulated environment.

1 Introduction

Mobile communications have expanded the horizons of signal processing in the modern era of communication technologies. Several aspects of digital and analog signal processing affect the performance of communication. To achieve benefits, the minimum expected performance has to be met by concerned organization or telecommunication companies or their signal processing logic. Nowadays, when the planet is connected all over, *mobility* has also become common to all. The issues related to mobility can weaken the performance of the signals being transmitted. In addition, in the technologically rapidly emerging world, new applications arrive at quick rate which everyone wants to avail. *Reconfigurability* is another factor associated in this concern. The target is that the incoming signal should reach with minimum possible attenuation to the destination. Mechanisms to smoothen the received signal have to be *portable* or ready to use. This can be beneficial not only to improve the performance but also to achieve better signal reception quality.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 113–121.
springerlink.com

In general terms, signal processing is done through hardware based radio consisting of the components as illustrated in Figure 1. The use of hardware circuitry limits the researchers to make any dynamic change frequently. This difficulty led to search some alternative way to accomplish the same task more fairly. This is how SDR came into existence. It brings the capabilities of radio functionalities with signal processing functionalities to achieve reconfigurability and portability.

Software Defined Radio - this term was coined by Joseph Mitola[13] in 1991. As described in [1], "a basic SDR system may consist of a personal computer equipped with a sound card, or other analog-to-digital converter, preceded by some form of RF front end." Most of the signal processing tasks are handed over to the general-purpose processor instead of utilizing special-purpose hardware, thereby producing a transceiver that can receive and transmit different radio protocols or waveforms based solely on the software used. SDR fecilitates as a single wireless device which supports a wide range of functionalities.

GNU Radio Project[12], founded by Eric Blossom is an open source software radio community that makes it possible to add reconfigurability to existing signal processing packages. GNU Radio can be considered as a signal processing toolbox which can be customized as per the need. Signal Processing blocks are written in C++ and mapped into Python using simplified wrapper and interface generator (SWIG). GNU Radio Companion (GRC) is a GUI which can be used for convenience in programming.

A signal, when transmitted over a channel, becomes corrupted and reaches its destination with some alteration. Unless the communication is critical, this alteration may be acceptable. In order to ensure correct delivery of data, mechanisms to reduce the noise have to be devised. Use of filters is one such method, but it requires techniques to approximate noise pattern in the signal. Huang et al.[6] proposed Empirical Mode Decomposition that decomposes a signal into different monotonic signals, commonly

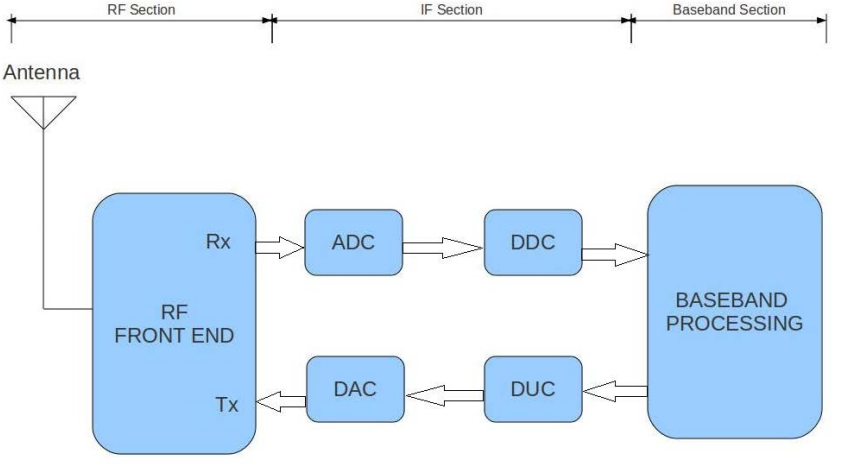


Fig. 1. Block Diagram for Hardware Defined Radio

known as intrinsic mode functions. This method is extremely helpful to separate a signal into several mono-component signals, which may be processed later on.

2 Motivation

Many researchers have thoroughly gone through the use of GNU Radio (an open source SDR) and contributed their findings in the area of wireless protocol testing or implementing various receivers. But a little research has been done in the area of writing a custom signal processing application for this tool. The challenge was to find a suitable way for noise reduction from an incoming signal, thus reducing the bit error rate (BER). It was found that till date, there haven't been any inbuilt signal processing block in GNU Radio to support this task except usual filtering blocks (which in turn, requires input parameters such as frequency or the pattern structure of the noise, which may not always be possible to estimate). Especially in wireless domain, one cannot predict the exact noise pattern in advance. Our work was centered around exploring or establishing some method that could help solve this problem. Empirical Mode Decomposition is useful to decompose the input signal into monotonic signals, which may later be useful to identify the noisy components in the signal.

Our target was to make it possible to utilize the capabilities of SDR to reduce the hardware based tasks and maximize software based computations so as to achieve reconfigurability. SDR, thus in real terms, enables *R*econfigurable *A*daptive *D*ynamic *I*nput *O*utput.

3 Empirical Mode Decomposition

Mathematical formalization of Empirical Mode Decomposition, as mentioned in [10], is described in Section 3.1.

3.1 Mathematical Concept

The Empirical Mode Decomposition (EMD) is an iterative process which decomposes real signal $f(t)$ into simpler signals (modes).

$$f(t) = \sum_{j=1}^{M} \phi_j(t) \tag{1}$$

Each monocomponent signal ϕ_j, with amplitude $r(t)$, should be representable in the form

$$\phi(t) = r(t)sin\theta(t) \tag{2}$$

These monocomponent signals ϕ, called Intrinsic Mode Functions (IMF), are produced by Empirical Mode Decomposition. EMD decomposes the signal into finite number of IMFs. Moreover, these IMFs reflect the intrinsic and reality information of the analyzed signal. Therefore, EMD method is a self-adaptive signal-processing method that is suitable for the analysis of non-linear and non-stationary process[7].

3.2 Intrinsic Mode Function

A function $\lambda(t)$ is defined to be an intrinsic mode function[10], of a real variable t, if it satisfies two characteristic properties:

1. λ has exactly one zero between any two consecutive local extrema.
2. λ has zero local mean.

Part A of Figure 2 shows several IMFs in increasing order generated after applying EMD on a speech segment. Part B shows corresponding FFTs of the produced IMFs. This shows that the IMF generation takes place in the decreasing order of their frequencies.

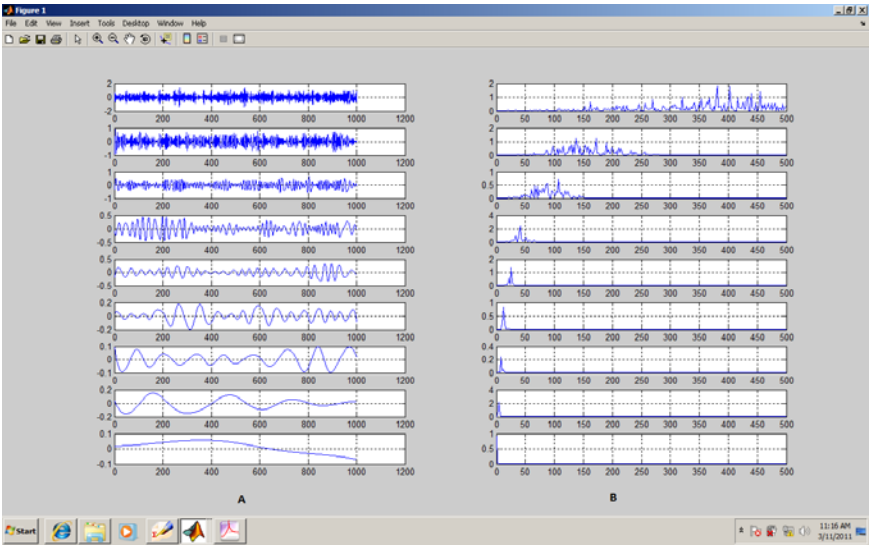


Fig. 2. Intrinsic Mode Functions and their Fourier transforms

The entire method for performing EMD is done through the sifting process. Cubic Spline interpolation is used to link local maxima and minima to form upper envelope and lower envelope of the signal respectively. The mean of these two envelopes is subtracted from the original signal. EMD is obtained after applying this process repeatedly. The sifting algorithm is highly adaptive; it is also unstable. A small change in data can often lead to different EMD[11].

4 Related Work

The results of [5] show that the IMF plots reveal that when noise is added to a clean speech signal, the first few IMFs contain most of noise energy and some of the speech. Our goal is to distinguish which IMF contain the speech or noise. This decomposition

pushes a significant amount of the speech energy to latter IMFs along with some residual noise. The reconstruction process is given in Equation 3, which involves combining the n IMFs and the residual $r[n]$.

$$x[n] = \sum_{i=1}^{n} IMF[n] + r[n] \tag{3}$$

Issues such as the stopping criterion for the sifting process or determining a specific spline interpolation for EMD are crucial for the efficiency of the algorithm. The results may vary due to highly adaptive nature of the sifting algorithm and ad hoc nature of using cubic splines. In [3], the cubic splines were replaced by B-splines, which gives an alternative way for EMD. But again this modification does not resolve those issues. The convergence problem has been addressed in [11] using iterating filters, but it provides similar results. In [7], it is shown that the IMFs defined by their energy difference tracking method meet the orthogonality condition and reflect the intrinsic and reality information of the analyzed signal.

The authors in [9] discuss a very good comparison of different assessing alternatives for the sifting process and introduce the use of rational splines which results into tradeoffs relative to the original cubic spline method. They are succeeded to reduce the over- and undershooting problems, but at the expense of more IMFs and more sifting.

In [4], the authors discuss the assets of the EMD as its sparseness and completeness. They also explain the weakness of this algorithm regarding its dependancy on the sifting convergence criterion or interpolation method.The authors describe in [8] the use of Empirical Mode Decomposition for denoising signals in an efficient manner. They suggest that it should be possible to separate out the noise portion from the incoming signal. A very good documentation in [2] is provided on how to write a custom block inside GNU Radio.

5 Scenario for Experiment

5.1 Initial Setup

A typical communication link includes, at a minimum, three key elements: a transmitter, a communication medium (or channel), and a receiver. The ability to simulate all these functions is required to successfully model any end-to-end communication system. As our goal was oriented towards improvement of the signal using Empirical Mode Decomposition, we used Matlab at the initial stage to test EMD and its effect over the noisy signals.

5.2 Simulation on Speech Signal

Since Empirical Mode Decomposition performs better with non-stationary signals, we chose speech signal for our simulation.

The simulation result of [5] shows that for low E_b/N_o, EMD method improves BER performance by approximately $3dB$ gain. These BER improvements can help solving the problem related to call drop outs and improves overall QoS.

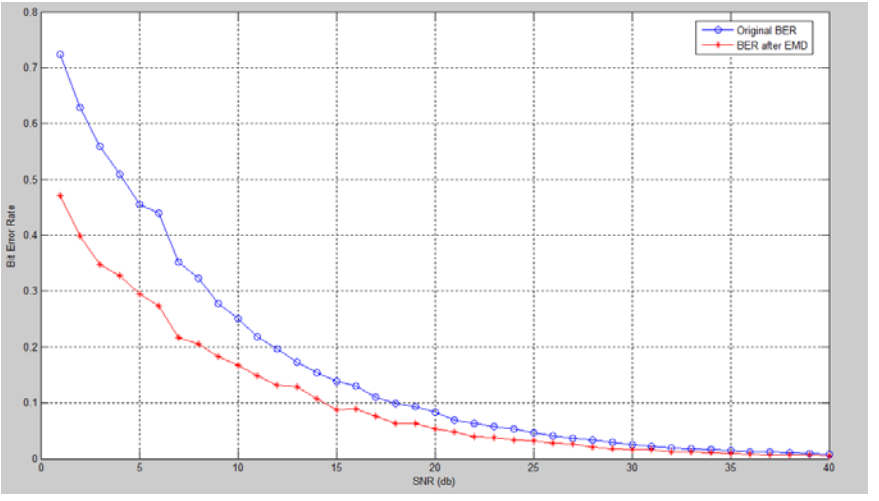


Fig. 3. Effect of EMD over Speech Signal

At the initial stage, we used a sample speech segment. The signal was passed through AWGN channel with varying SNR values. The EMD processed signal and unprocessed signal show significant difference. This is depicted using Figure 3.

5.3 Simulating GSM Signal

At present, GSM and CDMA are maximally used mobile communication signals. Our approach was to explore the effect of EMD as a denoising technique on different types of noise and embedding its functionalities into software defined radio.

At first, we constructed GSM signal by passing a random binary pattern through GMSK modulator, provided inside GNU Radio. This signal was corrupted with three different kinds of noise:gaussian noise, uniform noise and random noise. After applying EMD on this noisy signal, we compared the corresponding outputs produced after the experiment. It is observed that initial IMFs exhibit noise components in the signal more dominantly, hence our approach was to subtract the first few IMFs from the output of EMD. We started with removing only the first IMF from the output, which would contain the maximum noise energy. We refer to this approach as *Case I*.

We carried out simulation by varying different parameters such as type of noise, type of splines used in EMD and using first two IMFs for subtraction. Figure 4 describes the flow graph of the simulation. The analysis of results is shown in Table 1. It is clearly seen that EMD provides a good alternative to reduce bit error rate of the incoming signal.

5.4 Results

This simulation was handled for 2000 samples of GMSK signal and the results show that EMD performs better in presence of Gaussian Noise and it does not have any

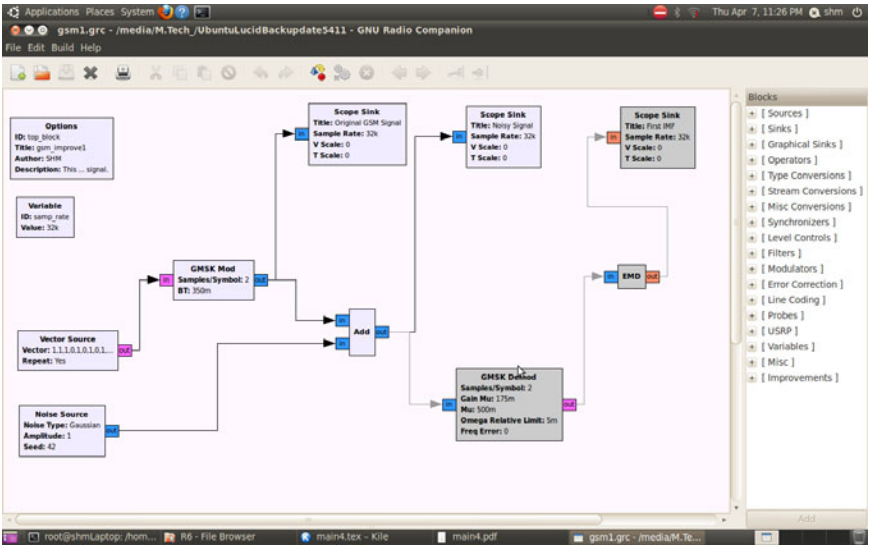


Fig. 4. Set up of Experiment in GNU Radio Companion

Table 1. Comparison of BER with and without EMD removing only the first IMF

Type Of Noise	Time(sec)	BER	BER with EMD
Gaussian	33.5	0.908	0.83
Uniform	19.5	0.914	0.869
Random	17	0.873	0.883

improvement during random noise. For GSM signals, it shows that for increasing SNR values, the BER remains the constant. After performing the simulation, we came to know that EMD processed signal is less noisy as compared to the unprocessed signal.

The underlying observation indicates that EMD should be avoided while random noise is active. We conducted another experiment with a little change in the scenario. We extracted first two IMFs from the noisy signal and deducted them from it. We refer to this approach as *Case II*. The results after this modification are depicted in Table 2.

The result after the modification shows that EMD has the same effect over Gaussian and Uniform noise, whereas there is a significant improvement over random noise in this case. This is because the most of the noise energy is concentrated in the initial IMFs. This may vary for different signals, but usually it may be contained within the first three IMFs.

The naive algorithm for EMD uses cubic natural spline. We implemented *Case II* approach using nearest neighbor spline and linear spline interpolation and their results are presented in Table 3 and Table 4 respectively. Results show that cubic interpolation is the most suitable approach for all the cases.

Table 2. Comparison of BER with and without EMD removing first two IMFs

Type Of Noise	Time(sec)	BER	BER with EMD
Gaussian	33.5	0.91	0.76
Uniform	19.2	0.914	0.734
Random	16.7	0.873	0.76

Table 3. Comparison with Nearest Neighbor Spline

Type Of Noise	Time(sec)	BER	BER with EMD
Gaussian	33.8	0.90	0.76
Uniform	19.4	0.92	0.74
Random	16.8	0.876	0.76

Table 4. Comparison with Linear Spline

Type Of Noise	Time(sec)	BER	BER with EMD
Gaussian	33.8	0.90	0.76
Uniform	19.2	0.92	0.74
Random	16.7	0.875	0.76

6 Conclusion

Empirical Mode Decomposition provides better results for signals having low signal to noise ratios. In addition, it is also observed that this algorithm works well when we remove the first two IMFs from the signal under process.

In presence of the uniform noise and gaussian noise, EMD exhibits about 20% improvement, whereas in case of random noise, for *Case I*, its performance is deteriorated by about 2%. Again, the cubic interpolation gives better result (about 1 to 2% improvement) over other methods.

Empirical Mode Decomposition provides a basis for separating out the noise dominated components from the signal. This is very useful in the data critical applications where maximal error free communication is expected.

7 Future Scope

Simulation of GSM signal for the experiment provides a good insight that wireless communication can be made noise prone at some extent using the method discussed in this work.

In this paper, we presented the advantage of empirical mode decomposition for denoising a signal. This work can be extended by implementing variants of EMD technique and providing more parameters to choose while performing the denoising using

EMD inside GNU Radio. Implementation can be more realistic if any hardware such as USRP is attached to GNU Radio so as to receive the OTA (over the air) files and process these live signals directly.

References

1. Bard, J., Kovarik V.J. : *Software Defined Radio- The Software Communications Architecture.* John Willey and Sons. (2007)
2. Blossom, Eric. How to write a signal processing block. Retrieved from www.gnu.org/software/gnuradio/doc/howto-write-a-block.html
3. Chen, Q., Huang, N., Riemenschneider, S. and Xu, Y. A B-spline approach for Empirical Mode Decompositions. *Advances in Computational Mathematics*, 24:171–195, 2006. 10.1007/s10444-004-7614-3.
4. Equis Sébastien and Jacquot Pierre. The empirical mode decomposition: a must-have tool in speckle interferometry? *Opt. Express*, 17(2):611–623, Jan 2009.
5. Guo, Z., Z. Xu, F. Wang and B. Huang,, 2010. Empirical Mode Decomposition for BER improvement in cellular network. Inform. Technol. J., 9: 146-151.
6. Huang, N.E., et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995, March 1998.
7. Junsheng, C., Dejie, Y., and Yu, Y. (2006). Research on the Intrinsic Mode Function (IMF) criterion in EMD method. Mechanical Systems and Signal Processing, 20(4), 817-824. Retrieved from http://linkinghub.elsevier.com/retrieve/pii/S0888327005001494
8. McLaughlin, S., Kopsinis, Y. : Empirical Mode Decomposition based denoising techniques.
9. Peel, M.C., G.G.S. Pegram and T.A. McMahon. : *Empirical Mode Decomposition : Improvement and Application.*
10. Sharpley, R.C., Vatchev, V. : *Analysis of the Intrinsic Mode Functions*. Industrial Mathematical Institute, Universitu of South Carolina.
11. Wang, Y., Lin L. and Zhou, H. Iterative filtering as an alternative algorithm for empirical mode decomposition. *Advances in Adaptive Data Analysis*, 1(4):543–560, 2009.
12. Official Gnu Radio Website, http://www.gnuradio.org.
13. Official Website of Joseph Mitola. http://web.it.kth.se/ maguire/jmitola/.

An Approach to Detect Hard Exudates Using Normalized Cut Image Segmentation Technique in Digital Retinal Fundus Image

Diptoneel Kayal and Sreeparna Banerjee

School of Computer Science & Engineering, West Bengal University of Technology, Kolkata, India
diptoneel@gmail.com, sreeparnab@hotmail.com

Abstract. Diabetic retinopathy is a disease commonly found in case of diabetes mellitus patients. This disease causes severe damage to retina and may lead to complete or partial visual loss. As changes occurs due to the disease is irreversible in nature, the disease must be detected in early stages to prevent visual loss. One of the most important sign of presence o f diabetic retinopathy in diabetes mellitus patients is the exudates. But detection of exudates in early stages of the disease is extremely difficult only by visual inspection. But an efficient automated computerized system can have the ability to detect the disease in very early stage. In this paper one such method is discussed.

Keyword: Diabetic retinopathy, exudates, median filtering, thresholding, Ncut.

1 Introduction

Computerized analysis of medical images is gaining popularity day by day, as it often produces higher sensitivity irrespective of experience of the analyst. For this reason efficient image analysis technique is used in a number of fields. The retina is the innermost and most important layer of eye, where the earliest pathological changes can be observed. It is composed of several important anatomical structures, which can indicate many diseases such as hypertension, diabetes and other various diseases of eye. The most effective way to detect these diseases is to regular screening of retinal fundus image. Diabetic retinopathy is one of the most serious complications of diabetes mellitus and a major cause of blindness. It is a progressive disease classified according to the presence of various clinical abnormalities. It is the most common cause of blindness among people aged 30-69 years [1]. One-fifth of patients of diabetes type II, have retinopathy at the time of diagnosis. In type I diabetes, diabetic retinopathy never occurs after diagnosis. But after 15 years all most of all patients with type I and two-third of those with type II diabetes have background of diabetic retinopathy [1]. In case of diabetic retinopathy blood vessels get damaged and protein and fat based particles gets leaked out of the damaged blood vessels beside blood flow to the retina decreases. These particles are referred to as exudates. Various methods have been developed for detection of exudates. These include thresholding and edge detection based techniques [2], FCM based approach [3], gray level

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 123–128.
springerlink.com

variation based approach [4], multilayer perceptron based approach [5]. Optic disk must be detected and segmented early in the detection process, as often optic disk has more or less same brightness and contrast as the exudates. So, if optic disk is not segmented in early stage the process may produce wrong result.

Thresholding and edge detection based approach [2] uses histogram specification and contrast enhancement procedure after segmentation of optic disk. Then dynamic thresholding is applied on the green channel and in final stage canny edge detector is used. Fuzzy c-means clustering (FCM) based approach [3] attempts to establish a dataset of candidate regions after preprocessing step which includes color normalization and contrast enhancement. A genetic algorithm and a multilayer neural network classifier are used to detect the exudates. As exudates have high gray level variation, this property is used in [4] to detect exudates. After establishing a dataset of possible exudate regions, morphological techniques are used to find out exact contours of the exudates. Multilayer perceptron based approach [5] attempts to extract a set of features from image regions and the subset which best discriminates between hard exudates and retinal background is selected. The selected subset is then used as input to the multilayer perceptron classifier to obtain the final segmentation of hard exudates. The proposed algorithm accepts input image in RGB format. As grayscale image comprises of m-by-n image matrix, whereas RGB image comprises of m-by-n-by-3 image matrix and it is easier to operate on an m-by-n matrix than on m-by-n-by-3 grayscale image is used for whole process. If the input is in RGB then the input image is converted into a grayscale image for the ease of operation.

2 Method

The proposed method has five steps.

- Median filtering
- Image subtraction
- Thresholding
- Application of normalized cut image segmentation method (Ncut)
- Image addition

2.1 Median Filtering

Digital image noise usually appears in the high frequency of the image spectrum. So, a low-pass digital filter may be used for noise removal. A non-linear low-pass filter can remove noise while preserving the edges. Such a filter is based on *data ordering*. Let x_i, i=1,2,….n be n observations, whose number n=2v + 1 is odd, can be ordered according to their magnitude as follows:

$$x_{(1)} < x_{(2)} < \ldots\ldots\ldots < x_{(n)} \tag{1}$$

x_i denoted i-th *order statistic*. $x_{(1)}$, $x_{(n)}$ are the maximum and minimum observations respectively. The observation $x_{(v+1)}$ lies in the middle and is called *median* of the observations. It is also denoted by med(x_i). By definition, the median lies in the middle of the observation data. It minimizes the L_1 norm:

$$\sum_{i=1}^{n} |x_i - med| \rightarrow \min \tag{2}$$

According to (2) the median is the maximum likelihood estimate of the location for the Laplacian distribution:

$$f(x) = \frac{1}{2} \exp(-|x|) \tag{3}$$

A two-dimensional median filter has the following definition

$$y(i,j)=\text{med}\{ x(i+r,j+s), \quad (r,s)\in A \quad (i,j)\in Z^2\} \tag{4}$$

Where $Z^2 = Z \times Z$ denotes the digital image plane. The set $A \subset Z^2$ defines the filter window. If the input image is of finite extent $N \times M$, $0 \leq i \leq N-1$, $0 \leq j \leq M-1$, definition (4) is valid only in the interior of the output image, that is, for those *i,j* for which

$$0 \leq i+r \leq N-1, \ 0 \leq j+s \leq M-1, \ (r,s) \in A \tag{5}$$

(5) is not valid at the border of the image. There are two approaches solve this problem. In the first one, the filter window A is truncated in such a way so that (5) is valid and definition (4) can be used again. In the second approach, the input sequence is appended with sufficient samples and (4) is applied for

$$0 \leq i \leq N-1, \ 0 \leq j \leq M-1$$

Median filter can remove additive white noise. They are very efficient in the removal of noise having long-tailed distribution. The median is a robust estimator of location also. Therefore a single outlier (e.g. impulse) can have no effect on its performance, even if its magnitude is very large or small. The median becomes unreliable only if more than 50% of the data are outlier. The robustness of median filter makes it very suitable for impulse noise filtering. This property of median filter can be used for than that of its neighboring region if we apply median filter on the input image (in grayscale format) we would obtain a filtered image in which the exudates are blurred to a great extent. Beside as the optic disk part has almost same brightness as that of exudates this part of retinal fundus image will also becomes blurred. Median filter has another interesting property of preserving sharpness of edges of the image. Preservation of edge information and its enhancement is a very important subjective feature of the performance of digital image filter. Median filters not only smoothes noise in homogenous image regions but tends to produce regions of constant intensity. All these properties make median filter ideal for this approach.

2.2 Image Subtraction

Next step of this approach is subtraction of median filtered image from input image (in case of the input image is in grayscale form) or subtraction of median filtered image from grayscale form on input image (in case of the input image is in RGB form). Image subtraction is used to find changes between two images of same scene. Mathematically image subtraction can be denoted as, c(m,n)=f(m,n)-g(m,n).

As mentioned earlier median filtering makes the brighter regions (i.e. exudates) into blur, hence the result of subtraction gives us the output in which only regions with high brightness and contrast can be observed. Subtraction is one of the most important step of this process as in this step the desired features of input retinal fundus image is extracted.

2.3 Thresholding

If in an image consists of light objects an a dark background, in such a way that object and background pixels have intensity values grouped into dominant modes, then we can extract light objects from background using thresholding operation. Then any point (x,y) in the image at which $f(x,y) > T$ is called an object point, where T is known as threshold parameter. Otherwise the point is called a background point. Thresholding operation can be defined as follows:

$$g(x,y) = \begin{cases} 1 & \text{if } f(x,y) > T \\ 0 & \text{if } f(x,y) \leq T \end{cases}$$

Thresholding operation has been used successfully in this process and is applied on the subtracted image. This operation gives us the exudate regions on a dark background. In this approach value of threshold parameter is calculated experimentally. No automatic method is used for calculating value of threshold parameter.

2.4 Normalized Cut Image Segmentation

Normalized cut multiscale image segmentation procedure has been used in order to segment a digital retinal fundus image efficiently. In case of multiscale image segmentation, segmentation is done with image structures over image scales. The set of points in an arbitrary feature space is presented as a weighted undirected graph G=(V,E). Nodes of the graph are the points in the feature space. An edge is formed between every pair of nodes and the weight on each edge W(i,j) is a function of similarity between nodes i and j. A graph G=(V,E) is partitioned into two disjoint complementary sets A and B, B=V-A, by removing the edges connecting two parts. The degree of dissimilarity between two sets can be computed as a total weight of removed edges. That closely relates to a mathematical formulation:

$$\text{cut}(A,B) = \Sigma_{u \in A,\, v \in B}\ w(u,v)$$

Shi and Malik [6] proposed a modified cost function of name normalized cut. Instead of looking at the value of total edge weight connecting two partitions, the proposed measure computes the cut cost as a fraction of total edge connection to all nodes.

$$\text{Ncut}(A,B) = \frac{cut\ (A,B)}{asso\ (A,V)} + \frac{cut\ (A,B)}{asso(B,V)}, \quad \text{asso}\ (A,V) = \Sigma_{u \in A,\, t \in V} w(u,t)$$

2.5 Image Addition

Image addition is used to create double exposure. If *f(m,n)* and *g(m,n)* represents two images then addition of these two images to get the resultant image is given by $c(m,n) = f(m,n) + g(m,n)$.

If multiple images of a given region are available for approximately the same data and if a part of one of the image has some noise then the part can be compensated from other images available through image addition.

3 Result

The proposed algorithm is implemented MATLAB 7.0.4. Configuration of the computers used for development and testing purpose is Intel Core2Duo 1.5 GHz processor and 1 GB of RAM. The algorithm is tested on a database of ten images, among which seven images with exudates and three images with no exudates. The ground data is verified by an expert ophthalmologist. For evaluation purpose value of True Positive (TP), False Positive (FP) and False Negative (FN) parameters are determined for each image. Sensitivity (S) and Predictivity (P) is used for the measurement of accuracy. Sensitivity and predictivity are defined as follows.

$$\text{Sensitivity (S)} = \frac{TP}{TP+FN}$$

$$\text{Predictivity (P)} = \frac{TP}{TP+FP}$$

The overall sensitivity and predictivity are found to be 97.9% and 95.91% respectively.

Table 1.

	TP	FP	FN	$S = \frac{TP}{TP+FN}$	$P = \frac{TP}{TP+FP}$
IMAGE 1	5	0	0	100	100
IMAGE 2	8	0	0	100	100
IMAGE 3	7	0	0	100	100
IMAGE 4	11	1	2	84.60	91.67
IMAGE 5	14	0	2	100	87.50
IMAGE 6	13	2	3	86.60	81.25
IMAGE 7	4	0	0	100	100
IMAGE 8	0	0	0	100	100
IMAGE 9	0	0	0	100	100
IMAGE 10	0	0	0	100	100
OVERALL				**97.12%**	**96.04%**

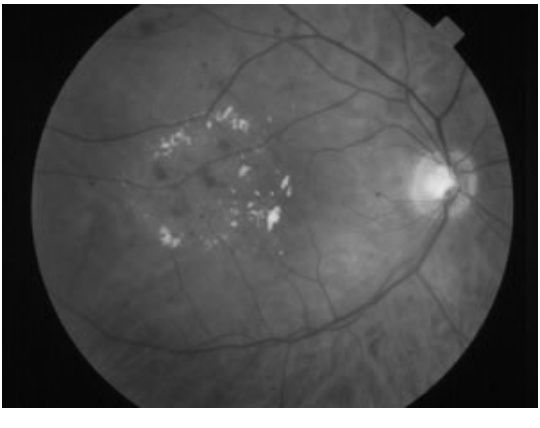


Fig. 1. Input image

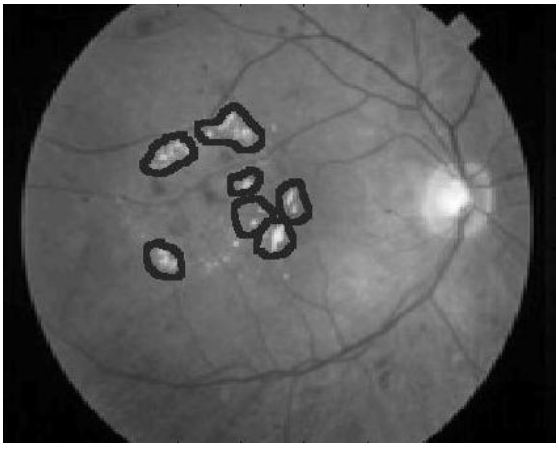


Fig. 2. Output image

References

[1] Klein, R., Klein, B., Moss, S., Davis, M., Demants, D.: The Wisconsin epidemiology study of diabetic retinopathy type II. Archieve of Opthalmology 102(4), 520–526 (1984)
[2] Sagar, A.V., Balasubramaniam, B., Chandrasekhara, V.: A Novel Intergrated Approach Using Dynamic Thresholding and Edge Detection for Automatic Detection of Exudates in Digital Fundus Retinal images. In: IEEE International Conference on Computing, pp. 286–292 (2007)
[3] Li, H., Chutatape, O.: A Model Based Approach for Automated Feature Extraction in fundus Images. In: IEEE International Conference on Computer Vision, pp. 127–133 (2003)
[4] Osrach, A., Shadgar, B., Markmham, R.: A Computational Intelligence Based Approach for Detection of Exudates in Diabetic Retinopathy. IEEE (2009)
[5] Walter, T., Klein, J.C., Massin, P., Erginay, A.: A contribution of Image Processing To the Diagnosis of Diabetic Retinopathy – Detection of Exudates in Color Fundus Image of Human Retina. IEEE Transaction on Medical Imaging 21(10), 256–264 (2002)
[6] Shi, J., Malik, J.: Normalized Cut Image Segmentation. In: International Conference on Vision and Pattern Recognition, San Juan, Puerto Rico (June 1997)
[7] Garcia, M., Hornero, R., Sanchez, C.I., Lopez, M.I., Diez, A.: Feature Extraction and Selection for Automatic Detection of Hard Exudates in Retinal Images. In: Conference of the IEEE EMBS, France (2007)
[8] Sinthanayothin, C., Kongbunkiat, V., Phoojaruenchanachai, S., Singalavanija, A.: Automated Screening System For Diabetic Retinopathy. In: 3rd International Symposium On Image And Signal Processing And Analysis, pp. 915–920 (2003)
[9] Ravishankar, S., Jain, A., Mittal, A.: Automated Feature Extraction for Early Detection of Diabetic Retinopathy in Fundus Images, pp. 210–218. IEEE (2009)
[10] Wareham, N.J.: Cost Effectiveness of Alternative Methods for Diabetic Retinopathy Screening. Diabetes Care 16, 844 (2003)
[11] Liu, Z., Opas, C., Krishnan, S.: Automatic Image Analysis for Fundus Image. In: 19th International Conference of IEEE EMBS, Chicago, pp. 524–528 (1997)
[12] Kayal, D., Banerjee, S.: A Simplified Method to Detect Hard Exudates in Digital Retinal Fundus Image. In: International Conference On Biomedical Engineering and Assistive Technologies, Jalandhar, India

Latency Study of Seizure Detection

Yusuf U. Khan[1], Omar Farooq[2], Priyanka Sharma[2], and Nidal Rafiuddin[1]

[1] Electrical Department, Aligarh Muslim University, Aligarh
yusufkhan.ee@amu.ac.in, nidal.rafi@gmail.com
[2] Electronics Department, Aligarh Muslim University, Aligarh
omarfarooq70@gmail.com, priya.32dec@gmail.com

Abstract. Epilepsy is a physical condition that occurs when there is a sudden, brief change in the normal working of brain. At this time, the brain cells are unable to function properly and the level of consciousness, movement etc. may get affected. These physical changes occur due to the hyper-synchronous firing of neurons within the brain. Most of the existing methods to analyze epilepsy depend on visual inspection of EEG recording of patients by experts who are very small in number. Also this method takes more time in diagnosis of epilepsy since EEG recording creates very lengthy data. This makes automatic seizure detection necessary. In this study a method to detect the onset of seizures is proposed in which the latency in detecting the onset has been decreased very much. The proposed method detected the onset of seizures with the mean latency of 0.70 seconds when applied on CHB-MIT scalp EEG database.

Keywords: Epilepsy, Seizures, EEG, Latency.

1 Introduction

Epilepsy, one of the most common neurological disorders, is characterized by recurrent seizures that may cause loss of consciousness or a whole body convulsion. The seizures are random in nature and patients are often unaware of the occurrence of them which may increase the risk of physical injury. Studies show that about 50 million people worldwide have been suffering from this disease [2]. For the treatment of epilepsy, patients take antiepileptic drugs (AEDs) on a daily basis but unfortunately despite treatment about 25% of the patients continue to experience frequent seizures [11]. These patients suffer from the epilepsy that does not respond to AED and called as refractory epilepsy. Surgery is the most effective and generally adopted treatment for these patients, but can be done only when epileptogenic focus is identified accurately. For this purpose different type of tracers are employed as soon as possible after onset detection. Early detection of seizure onset would be helpful in the rapid injection of tracer and hence accurate localization of epileptogenic focus.

EEG has been an important clinical tool for the analysis and treatment of epilepsy [12]. The EEG is a multichannel recording that reflect the activity generated by number of neurons within the brain. It is generally recorded using the electrodes placed on the scalp. Visual inspection of the EEG data is done by specialists to

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 129–136.
springerlink.com

analyze epilepsy. But observing EEG continuously for a long time is a very tedious task, since EEG data recordings create lengthy data [4]. Hence automatic seizure detection is essential in clinical practice

Automatic detection of seizures through the analysis of scalp EEG has been an important area of research for the last few decades. In 1976, Gotman and Gloor [6] proposed a method of recognition and quantification of interictal epileptic activity (spikes and sharp waves) in human scalp EEG. To perform the automatic recognition, the EEG of each channel was broken down into half waves. A wave was characterized by the durations and amplitudes of its two component half waves, by the second derivative at its apex measured relative to the background activity, and by the duration and amplitude of the following half wave. This method gave a good basis to the work in the field of seizure detection. The main limitation of the method was the absence of precise definition for an interictal epileptic event. In 1982, Gotman [5] proposed an improved method for automatic detection of seizures in EEG. After this many methods have been proposed to detect the seizures, but few of those were on onset detection of seizures.

Qu and Gotman [8] proposed a patient specific seizure onset detection method and achieved a sensitivity of 100% with mean latency of 9.4 seconds. The average false detections declared were 0.02 per hour. The algorithm was tested on 47 seizures of 12 patients. The drawback of this method was the need of template for the detection of seizures. In 2004, Gotman and Saab [9] designed an onset detection system. When it was tested using scalp EEG of 16 patients having 69 seizures, sensitivity of 77.9% with false detection rate of 0.9 per hour and median detection delay of 9.8 seconds were reported. Sorensen et al [11] used matching pursuit algorithm and achieved 78-100% sensitivity with 5-18 seconds delay in seizure onset detection while at the same time 0.2- 5.3 false positives per hour were declared. The method was evaluated using both scalp and intracranial EEG. Shoeb and Guttag [10] reported 96% sensitivity and mean detection delay of 4.6 seconds when worked on CHB-MIT database [13].

In this paper, a method to study latency of seizure detection using wavelet based features and statistical features have been proposed. Daubechies wavelet has been widely used for the seizure detection in EEG [1, 3, 9]. The proposed algorithm uses the Daubechies wavelet (of order 4) to detect the onset of seizures present in the database.

2 Methodology

2.1 Dataset

The database used in this study was CHB-MIT scalp EEG database which is freely available online [13]. It was collected at the Children's Hospital Boston and consists of EEG recordings from pediatric subjects, suffering from intractable seizures. Recordings, grouped into 24 cases, were collected from 23 subjects (5 males, ages 3-22, 17 females, ages 1.5-19 and 1 unknown).

All EEG signals were sampled at 256 Hz with 16-bit resolution. Most files contain 23 EEG channels (24 or 26 in a few cases). EEG data was recorded according to the standard 10-20 system. Overall this 24 patient dataset consisted of 916 hours of continuously recorded EEG and 198 seizures. First 10 patient's EEG from this database was used for this study. The line frequency of 60 Hz was removed from the database.

2.2 Feature Extraction

Feature extraction is a crucial step of seizure detection in which features of the data are investigated that is able to differentiate between the seizure and normal EEG data. In this study four features: relative energy, relative σ^2/μ_a (ratio of variance (σ^2) and absolute mean (μ_a), also known as coefficient of dispersion), IQR (Interquartile range) and MAD (Mean absolute deviation) were extracted from the data. The method of computation of relative features is described later. The seizure data was divided into frames of 1 second each using non-overlapping epoch window. A background window of 25 seconds was taken to normalize the epoch features and this window was made to move with epoch window (Fig. 1a & 1b). A gap of 15 seconds between epoch & background was taken to prevent seizure onset into the background (Fig. 1a).

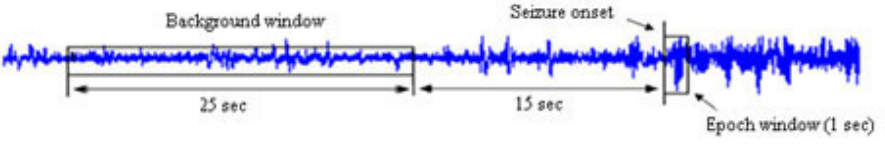


Fig. 1a. A segment of EEG showing seizure onset, background window and epoch window

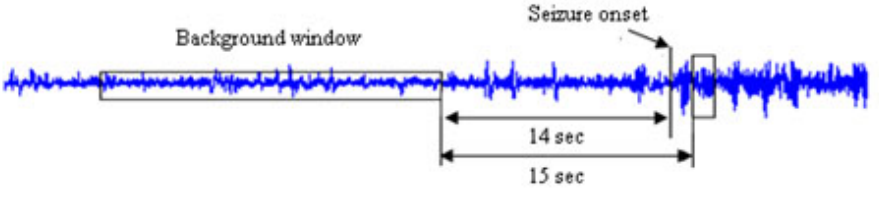


Fig. 1b. A segment of EEG, showing movement of background window

The background window was also divided into frames of 1 second and then each epoch of seizure & frames of background window were decomposed at level 5 using Daubechies wavelet of order 4. Since most of the seizure information lies in 0.5-30 Hz range, levels A5 (0-4 Hz), D5 (4-8 Hz), D4 (8-16 Hz) and D3 (16-32 Hz) were used for the computation of features. The used wavelet levels are shown in blue boxes in Fig. 2.

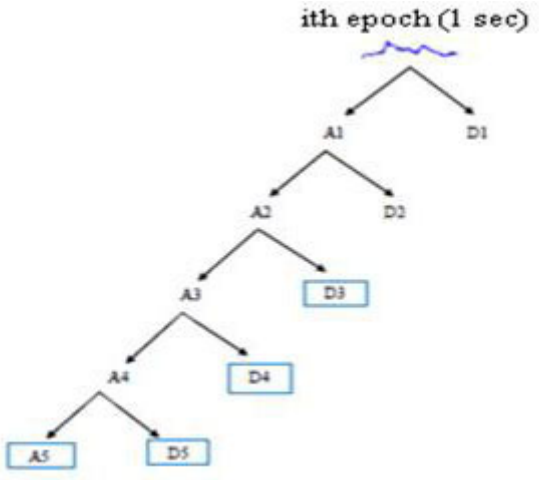


Fig. 2. Five level wavelet decomposition

Relative energy (E) for each epoch was computed using equation (1)

$$E = \frac{E_e}{E_b} \tag{1}$$

where, E_e is energy in epoch window and E_b is average energy in background window. These can be calculated using equations (2) & (3) respectively.

$$E_e = \sum_{n=1}^{256} x^2(n) \tag{2}$$

$$E_b = \frac{1}{25}\sum_{k=1}^{25}\sum_{l=1}^{256} x^2(l) \tag{3}$$

where, x is the sample value, n is the time sample of epoch, k is the frame no. of background window and l is the time sample of each frame of background window.

Next, variance and absolute mean of the epoch were computed. In addition, variance and absolute mean of each frame of background window to normalize epoch features were evaluated. Then above methodology was repeated to calculate relative σ^2/μ_a.

IQR and MAD were the statistical features computed over the raw EEG. Consequently total 10 features were extracted for each epoch on a channel which is shown in Fig. 3.

Mostly 23 channels were used for recording in database. So for each epoch a vector of 23*10 dimensions was formed and because a seizure is of different duration (T), T such epochs would be there. A feature vector was formed by concatenating 23*10 dimension vector of each epoch vertically. Hence the dimension of feature

vector was (23*T)*10. This feature vector is for seizure EEG signal. Similarly feature vectors for normal EEG signal were calculated. Normal EEG signal of T sec duration was taken randomly from non-seizure records. It was assumed that normal data, used for feature vector formation, was free from artifacts. Since, 55 seizures were present in the first 10 patient's EEG of database, 55 feature vectors were formed for seizure and 55 feature vectors were similarly formed for normal EEG signals.

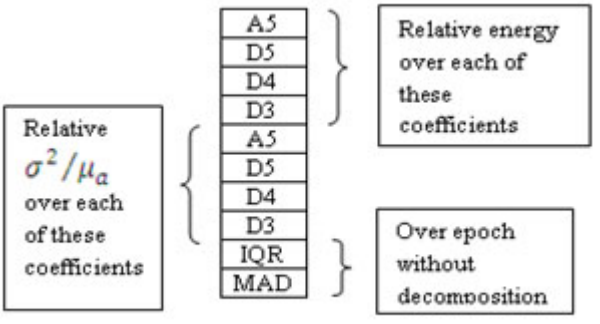


Fig. 3. Features used in this study

2.3 Results and Discussion

Classification between the normal and seizure EEG signal was done by inputting the extracted features to the linear classifier. These are of discriminative type i.e. they learn the way of discriminating the classes in order to classify a feature vector. It uses hyper-planes to separate the data representing different classes. If the problem is a two class problem such as seizure and non-seizure type, the class of feature vector depends on which side of hyper plane it lays. The separating hyper plane is that plane for which the distance between two classes' means is maximum and interclass variance is minimum [7]. Here the classification was done for each patient separately and the results obtained were averaged out to get the final result. For example patient 6 was having 10 seizures, so 8 of them were used for training and 2 were used for testing at a time and this process was repeated until every seizure got tested.

The classification was done to differentiate between two classes: seizure and normal EEG. Seizure epoch was labeled using 1 and normal epoch was labeled using 0. The classifier declared the seizure in any epoch if it was present in at least 40% channels. This was done to eliminate the artifact detection as seizures.

The performance of the classification was measured using the metrics: latency, sensitivity and false detection rate. Latency is the term used for the delay between the expert marked seizure onset and the detected seizure onset. Sensitivity refers to the number of seizures detected. False detection rate refers to the number of times the detector declared the seizure during the course of 1 hour when it was not present actually.

The mean latency with which the seizure declared the onset of every seizure was 0.70 seconds. Fig. 4 is showing the mean latency of each of the 10 patients.

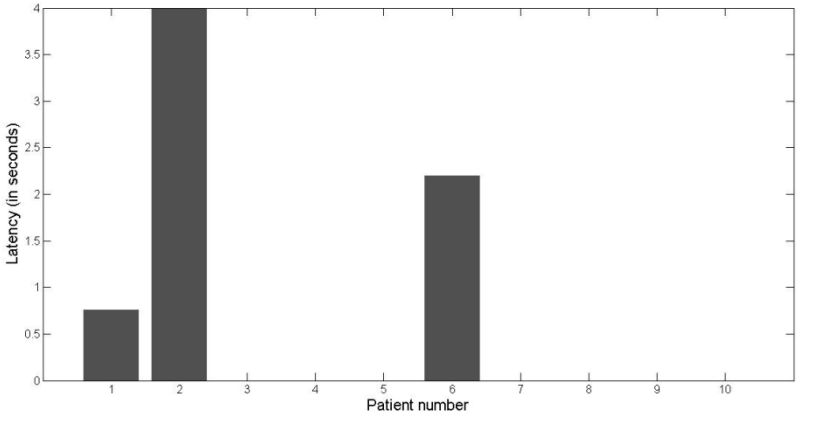


Fig. 4. Average latency for each patient

The maximum latency in seizure onset detection was observed in patient 2. Fig. 5a and 5b shows 10 sec EEG of one of the seizures present in patient 2 and background window taken for normalizing it respectively. The detector delays in detecting the onset of this seizure since its starting high amplitude characteristics are similar to background EEG characteristics.

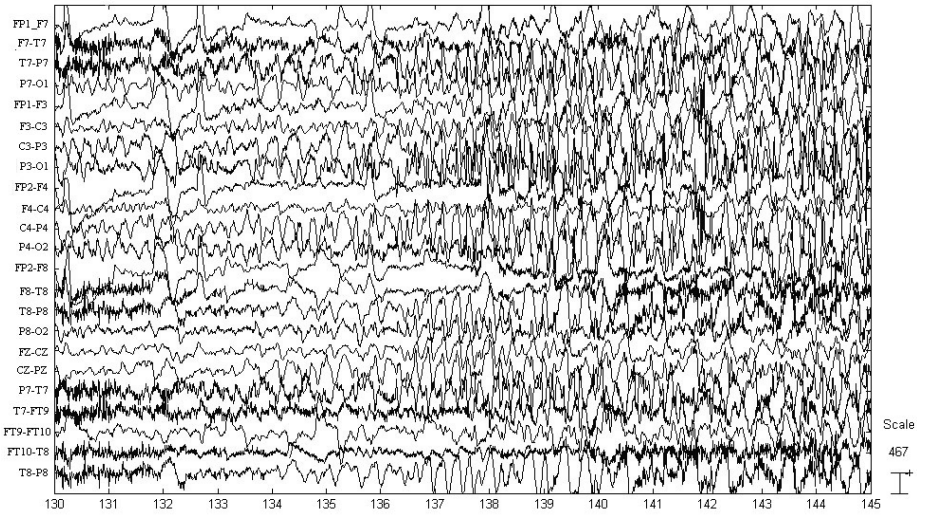


Fig. 5a. 15 seconds' EEG section of one of the seizures present in patient2

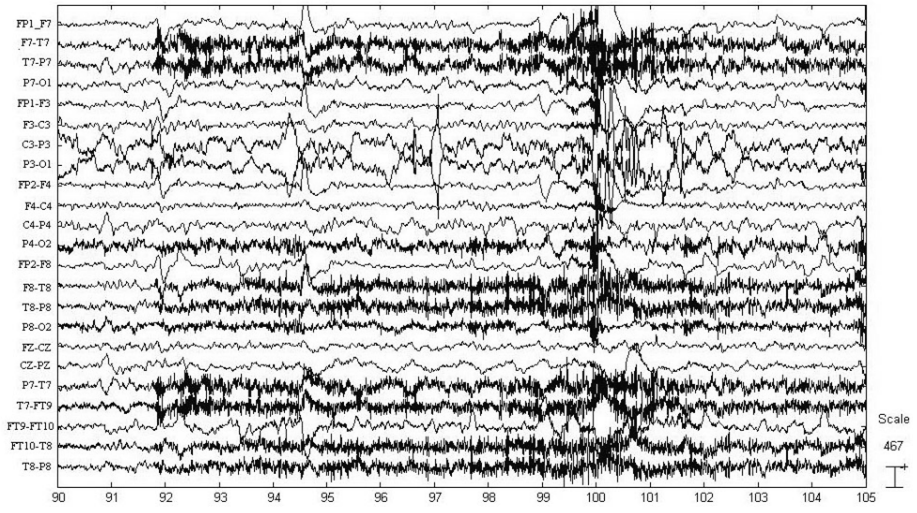


Fig. 5b. Background EEG taken to normalize the seizure shown in Fig. 5a.

Since the detector detected the onset of every seizure used for this study, hence the sensitivity achieved was 100%. The average number of false detections observed by the detector was 0.55 per hour. On comparison of the proposed method with the other's method who worked on the same database, the results observed are shown in Table 1. This comparison is done by considering only 10 patient's results from the results observed for 23 patients by Shoeb and Guttag [10].

Table 1. Comparison of the performance of our method with other's method

Performance metrics	Average Latency	Average Sensitivity	Average False Detection Rate
Our work	0.7 seconds	100%	0.55 false detections per hour
Shoeb and Guttag's method [10]	3.4 seconds	96.2%	0.10 false detections per hour

3 Conclusion

In this work, statistical features in combination with wavelet based features were extracted and a method to detect the onset of seizures with low latency has been proposed. The methodology was able to detect all the seizures (55 in 10 patients) while Shoeb and Guttag [10] reported a detection rate of 96.2% on the same data. In addition, we have obtained much lower latency (0.7 seconds) as compared to the latency achieved by [10] (3.4 seconds). However, the average specificity reported by Shoeb and Guttag [10] was better than our result which is acceptable when compared with the better sensitivity (100% in our case) achieved in this work. In future, we intend to add spatial and temporal context in order to improve the specificity. More features with additional discriminatory information will be investigated to further improve the results.

References

1. Adeli, H., Zhou, Z., Dadmehr, N.: Analysis of EEG records in an epileptic patient using wavelet transform. J. Neurosci. Methods 123(1), 69–87 (2003)
2. Dorai, A., Ponnambalam, K.: Automated epileptic seizure onset detection. In: 2010 International Conference on Autonomous and Intelligent Systems (AIS), pp. 1–4 (June 2010)
3. Fathima, T., Bedeeuzzaman, M., Farooq, O., Khan, Y.U.: Wavelet Based Features for Epileptic Seizure Detection. MES Journal of Technology and Management 2(1), 108–112 (2011) ISSN 0976-3724
4. Geetha, G., Geethalakshmi, S.N.: Detecting Epileptic Seizures Using Electroencephalogram: A New and Optimized Method for Seizure Classification using Hybrid Extreme Learning Machine. In: 2011 International Conference on Process Automation, Control and Computing (PACC), pp. 1–6 (July 2011)
5. Gotman, J.: Automatic recognition of epileptic seizures in the EEG. Electroencephalography and Clinical Neurophysiology 54(5), 530–540 (1982)
6. Gotman, J., Gloor, P.: Automatic recognition and quantification of interictal epileptic activity in the human scalp EEG. Electroencephalography and Clinical Neurophysiology 41, 513–529 (1976)
7. Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B.: A review of classification algorithms for EEG-based brain–computer interfaces. J. Neural Eng. 4, 1–13 (2007), http://iopscience.iop.org/17412552/4/2/R01
8. Qu, H., Gotman, J.: A patient-specific algorithm for the detection of seizure onset in long-term EEG monitoring: Possible use as a warning device. IEEE Transactions on Biomedical Engineering 44(2), 115–122 (1997)
9. Saab, M.E., Gotman, J.: A system to detect the onset of epileptic seizures in scalp EEG. Clinical Neurophysiology 16(2), 427–442 (2005)
10. Shoeb, A., Guttag, J.: Application of Machine Learning To Epileptic Seizure Detection. In: Proceedings of the 27th International Conference on Machine Learning, pp. 975–982. Omnipress, Haifa (2010)
11. Sorensen, T.L., Olsen, U.L., Conradsen, I., Henriksen, J., Kjaer, T.W., Thomsen, C.E., Sorensen, H.B.D.: Automatic epileptic seizure onset detection using Matching Pursuit: A case study. In: 2010 Annual International Conference of the Engineering in Medicine and Biology Society (EMBC), pp. 3277–3280. IEEE (2010)
12. Yaylali, I., Kocak, H., Jayakar, P.: Detection of seizures from small samples using nonlinear dynamic system theory. IEEE Transactions on Biomedical Engineering 43(7), 743–751 (1996)
13. CHB-MIT scalp EEG database, http://physionet.org/physiobank/database/chbmit/

Analysis of Equal and Unequal Transition Time Effects on Power Dissipation in Coupled VLSI Interconnects

Devendra Kumar Sharma[1], Brajesh Kumar Kaushik[2], and Richa K. Sharma[3]

[1] Department of ECE, Meerut Institute of Engineering and Technology, Meerut, UP India
[2] Dept. of Electronics and Computer Engg, Indian Institute of Technology, Roorkee, India
[3] Department of ECE, National Institute of Technology, Kurukshetra, Haryana, India
d_k_s1970@yahoo.co.in, bkk23fec@iitr.ernet.in,
mail2drrks@gmail.com

Abstract. Packing more circuits on chip is achieved through aggressive device scaling and/or increase in chip size. This results in complex geometry of interconnect wires on-chip. High density chips have introduced problems like interconnect noise and power dissipation. The CMOS technology is best known for making ICs owing to its low power static dissipation and ease of integration when compared to BJT technology. As the CMOS technology moved below sub micron levels, the power consumption per unit chip area has increased manifolds. Low power consumption leads longer battery life and lesser heat generation in the circuit. The overall performance of a chip is largely dependent on interconnects. This paper addresses the impact of equal and unequal transition time of inputs on power consumption in coupled interconnects. To demonstrate the effects, a model of two distributed *RLC* lines coupled inductively and capacitively is considered. Each interconnect line is 4 mm long and terminated by capacitive load of 30 fF. The power dissipation is measured for dynamically switching inputs.

Keywords: Equal/ Unequal rise time, Power dissipation, Dynamic switching.

1 Introduction

The continuous improvement in density has been mainly achieved by scaling down the device dimension. Sustaining the trend (Moore's law) has been fuelled by the abilities of designers to put more transistors on-chip. Scaling device dimensions below 0.2 µm has resulted large consumption of chip area for complex interconnections. Thus, with technology advancement, on-chip interconnects have turned out to be more and more important than transistor resource [1], [2]. As per international technology roadmap for semiconductors (ITRS) [3], the gap between interconnection delay and gate delay will increase to 9:1 and on-chip wire length is expected to increase to 2.22 km/cm^2 for future nanometer scale integrated circuits. So, for high speed high density chips, the chip performance is mostly affected by the interconnections rather than device performance.

The decrease in interconnect width and thickness leads to increase in resistance while short spacing between them progressively increases the parasitic capacitance.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 137–144.
springerlink.com

With high clock speed, faster signal rise time and longer wire lengths, the inductance of interconnect significantly plays a major role in on-chip circuit performance [4]. Due to the presence of these line parasitics effects, the *RLC* distributed model or transmission line model [2], [5] is more effective in current technology. The short spacing between interconnect wires, long wire lengths and high operational frequency causes significant value of coupling parasitics i.e. mutual inductance (M) and coupling capacitance (C_C) in the circuit. The resulting effect of these parasitics is crosstalk noise, propagation delay and power dissipation. These performance parameters affect the signal integrity.

The denser designs of integrated circuits introduce problems of power dissipation. So, the power analysis of integrated circuits becomes an important issue of study. Small power consumption makes the circuit/device more reliable. The CMOS circuits are best known for its low power consumption. Below submicron technology, the power consumption per unit area of CMOS chip has risen tremendously.

Broadly, the power dissipation in CMOS circuit consists of two factors i.e. static dissipation and dynamic dissipation. The static power dissipation contribute a small percentage to the total power dissipated in the circuit, where as the dynamic power dissipation, which is due to charging and discharging of parasitic capacitive load of interconnects and devices, contributes dominantly. The dynamic power dissipation may be written as

$$P = K\, C_T\, V_{dd}^2 f. \tag{1}$$

In Eq.(1), C_T comprises of interconnect parasitic capacitance and the load capacitance (C_L). The V_{dd} is power supply voltage, f is operating frequency and K is switching activity factor. There are two factors affective the dynamic power dissipation viz. clock speed and transistor density. If the transistors are switching more rapidly, the power dissipation will be more. The dynamic power dissipation is due to switching current and through current. The through current is due to short circuit path between supply and ground rails during switching. The power dissipated due to though current is always substantially smaller than switching power.

A great deal of research has been done on the analysis of crosstalk noise and delay [6], [7], [8], [9], [10], [11]. However, researches have reported the power dissipation in interconnects considering different aspects. Power related issues of CMOS in nanometer regime are reported in [12]. The analysis of power consumption in optimally buffered single line RC model is reported in [13]. In [13], the authors have described a method for power distribution analysis using reduced order model. The power dissipation is analyzed against variations in interconnect width for uncoupled line [14]. Power dissipation is analyzed against variations in load for capacitively coupled interconnects [15]. A method for analyzing power distribution using reduced order model is reported in [16]. This paper addresses the power dissipation dependencies on equal and unequal transition time of inputs in capacitively and inductively coupled interconnects. This study is equally important because of the fact that the two inputs may have different transition time due to different length of interconnects. The SPICE simulations are run and various waveforms are obtained .

This paper is organized in four sections. Section 2 describes the setup for simulation of interconnect model. The effect of equal and unequal transition time of inputs to two interconnect lines on power dissipation are observed and discussed in section 3. Finally, section 4 concludes this paper.

2 Simulation Setup

To demonstrate the effects of equal and unequal transition time of inputs on power dissipation, two uniformly distributed *RLC* lines coupled capacitively and inductively as shown in Fig. 1. are simulated.

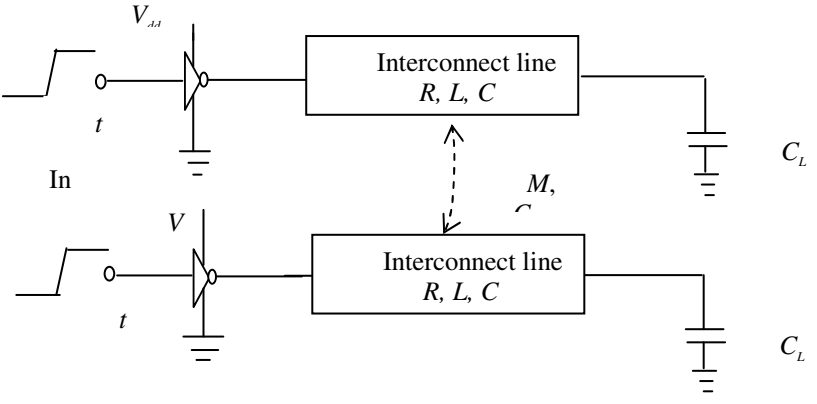


Fig. 1. Coupled interconnect lines

$$R = \begin{bmatrix} 12{,}500 & 0 \\ 0 & 12{,}500 \end{bmatrix} \quad C = \begin{bmatrix} 190p & -64p \\ -64p & 190p \end{bmatrix} \quad L = \begin{bmatrix} 1.722\mu & 1.4\mu \\ 1.4\mu & 1.722\mu \end{bmatrix}$$

Fig. 2. Interconnect parasitics

For our study, global interconnect is considered and it is assumed that there are several metal layers available for the interconnects. The length of each interconnect is taken as 4mm and each line of coupled structure is 2 µm wide, 0.68 µm thick and separated by 0.24 µm [11]. It is well accepted that the simulations of a distributed RLC interconnect line matches more accurately the actual behavior as compared to lumped model [2]. So, fifty distributed lumps of gamma type are taken for the length of interconnect under consideration. The parasitics values are obtained from expressions reported in [17], [18]. The far end of interconnect lines are terminated by a capacitive load of 30 fF. The interconnect parasitics matrices for one meter length are shown in Fig. 2. The simulations use an 1BM 0.13 µm technology node with copper interconnect process (MOSIS) with a power supply voltage of 1.5 V. The width of driver PMOS and NMOS are taken as 70 µm and 35 µm respectively.

3 Impact of Equal and Unequal Transition Time of Inputs

The transition time is defined as the time for a signal to go from 10 % to 90 % of its final value. The power dissipation is dependent on the supply current which in turn is

sensitive to input transition time. In this section, the impact of equal rise time ($tr_1 = tr_2$) and unequal rise time ($tr_1 \neq tr_2$) of inputs to interconnect lines are observed. The unequal rise time is because of different length which maps into different parasitic values of interconnects.

For our study, the rise time of both signals is varied equally from 10 to 760 ps. In case of unequal rise time of inputs, the difference i.e. $\Delta tr = tr_1 \sim tr_2$ is varied from 0 to 750 ps. The interconnect model under consideration is SPICE simulated for equal and unequal transition time of inputs. Both the cases of simultaneous switching of inputs are taken into consideration i.e.

Case I: Both inputs are switching in same phase i. e. from high to low or from low to high.
Case II: Both inputs are switching in opposite phase i. e. aggressor input in switching from high to low and victim input is switching from low to high.

3.1 Results and Observations

The impact of equal and unequal rise time of inputs on power dissipation in the model (Fig. 1) is shown in Fig. 3 to Fig. 6.

From these figures, following observations are drawn.

(i) From Fig. 3, if we compare the two modes of switching from high to low and low to high, it is observed that there is substantial difference in the value of power dissipation for transition time variation between 160 ps – 300 ps. However, the difference is at its peak at transition time of 160 ps. The peak value of this difference is 13.98 mW. Furthermore, it is observed from Fig 3 that power dissipation decreases monotonically from its peak value in case of high to low switching of inputs.

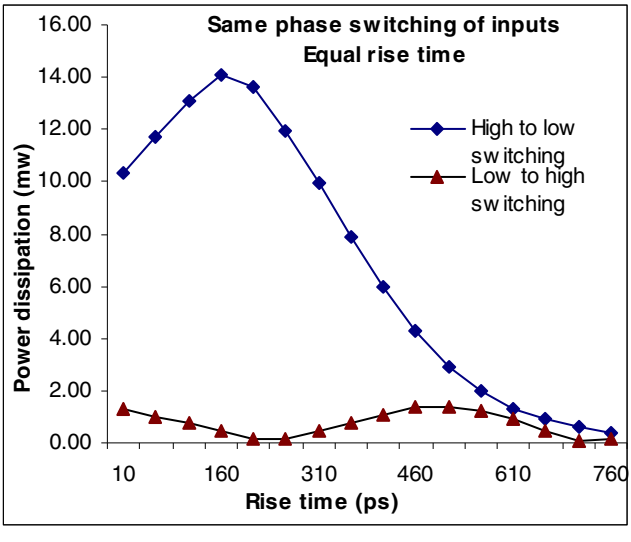


Fig. 3. Rise time as function of power dissipation in case of same phase switching and equal rise time of inputs

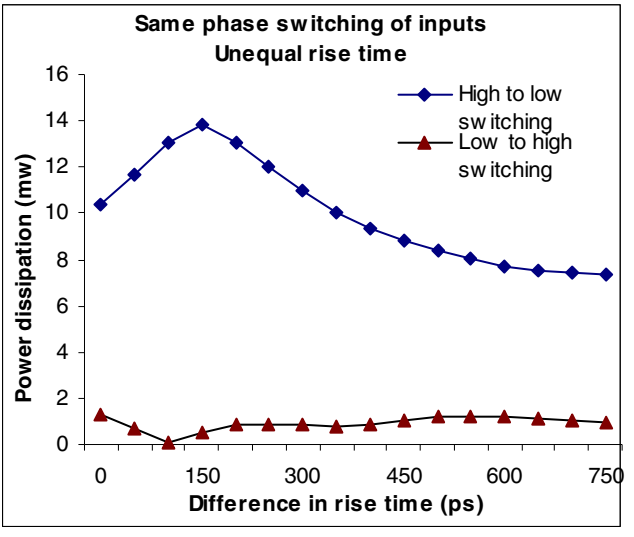


Fig. 4. Rise time as function of power dissipation in case of same phase. Switching of inputs and unequal rise time

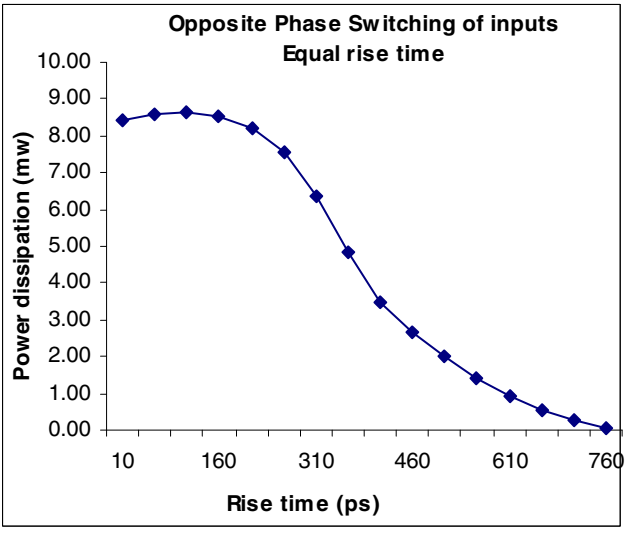


Fig. 5. Rise time as function of power dissipation in case of opposite phase switching and equal rise time of inputs

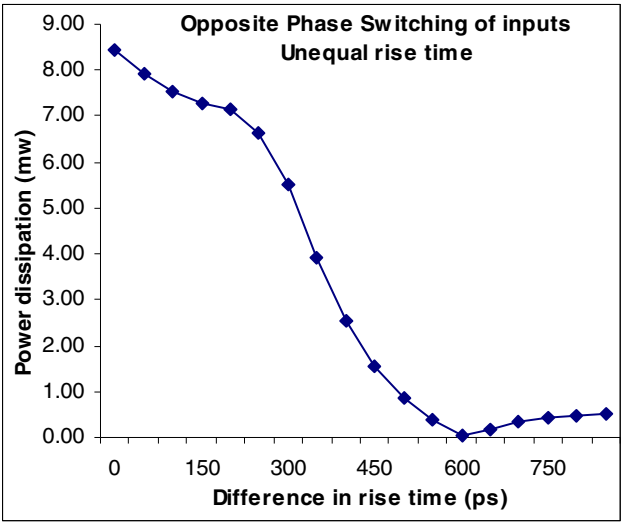


Fig. 6. Rise time as function of power dissipation in case of opposite phase switching and unequal rise time of inputs

(ii) Fig. 4 shows the simulation results when difference in rise time between inputs is varied for in- phase switching taking into consideration both the transitions from high to low and low to high. It is observed that, for inputs switching from high to low, the power dissipation first increases for low difference in the rise time of inputs and then starts decreasing monotonically. However, the power dissipation when inputs are switching from low to high is quite low as compared to high to low transition for whole range of difference in rise time under consideration. Furthermore, it is observed that the maximum difference in power dissipation is 13.78 mW between the two modes of switching from high to low and low to high.

(iii) From Fig. 5, it is observed that, in case of opposite phase switching of inputs when both the inputs are having equal rise time, the power dissipation decreases with increase in rise time. Furthermore, the maximum power dissipation of 8.66 mW is observed which is reasonably small than the value of maximum power dissipation observed in case of same phase switching of inputs (high to low) which is 14.10 mW. However, at large values of rise time in the specified range, the power dissipation in both the cases of simultaneous switching (Case I and Case II) is quite low.

(iv) From Fig.6, as the difference in rise time between inputs is increased, the power dissipation decreases dramatically and a minima of value 0.0423 mW occurs at 600 ps difference in rise time between inputs. If we compare these results with the results obtained in case of in- phase switching of inputs from low to high, the minima of value 0.055 mW occurs earlier i.e. at 100 ps difference in rise time between inputs.

4 Conclusion

This paper addressed the power dissipation issues in inductively and capacitively coupled VLSI interconnects for simultaneously switching inputs. In Case I when inputs are switching in same phase, the results are obtained for transition of inputs from low to high and high to low. From the results, it is concluded that the power consumption is low at large values of rise time in the range under consideration for both in phase and opposite phase switching of inputs. Furthermore, the impact of difference in rise time of the two inputs to interconnect lines on power consumption in the circuit is presented. It is observed that at reasonably large difference of rise time between inputs in the specified range, the power dissipation is low for both the cases of simultaneous switching.

References

1. Bakoglu, H.B.: Circuits, Interconnections and Packaging for VLSI, pp. 81–133. Addison-Wesley (1990)
2. Rabaey, J.M.: Digital Integrated Circuits: A Design Perspective. Prentice-Hall, Englewood Cliffs (1996)
3. Semiconductors Industry Association: International Technology Roadmap for Semiconductors, http://public.itrs.net
4. Ismail, Y.I., Friedman, E.G.: Figures of Merit to Characterize the Importance of On-Chip Inductance. IEEE Trans. on VLSI Systems 7(4), 442–449 (1999)
5. Ismail, Y.I.: On-chip Inductance Cons and Pros. IEEE Trans. on Very Large Scale Integration Systems 10(6), 685–694 (2002)
6. Kaushik, B.K., Sarkar, S., Agarwal, R.P., Joshi, R.C.: Crosstalk Analysis and Repeater Insertion in Crosstalk Aware Coupled VLSI Interconnects. Microelectronics International 23(3), 55–63 (2006)
7. Elgamel, M.A., Bayoumi, M.A.: Interconnect Noise Analysis and Optimization in Deep Submicron Technology. IEEE Circuits and Systems Magazine, Fourth quarter, 6–17 (2003)
8. Roy, A., Xu, J., Chowdhury, M.H.: Analysis of the Impacts of Signal Slew and Skew on the Behavior of Coupled RLC Interconnects for Different Switching Patterns. IEEE Trans. on VLSI Systems 18(2) (2010)
9. Chowdhury, M.H., Ismail, Y.I., Kashyap, C.V., Krauter, B.L.: Performance Analysis of Deep Submicron VLSI Circuits in the Presence of Self and Mutual Inductance. In: Proc. IEEE Int. Symp. Cir. and Syst., vol. 4, pp. 197–200 (2002)
10. Kahng, A.B., Muddu, S., Vidhani, D.: Noise and Delay Uncertainty Studies for Coupled RC Interconnects. In: Proc. IEEE Intl. Conf. on VLSI Design, pp. 431–436 (2004)
11. Sharma, D.K., Kaushik, B.K., Sharma, R.K.: Effect of Mutual Inductance and Coupling Capacitance on Propagation Delay and Peak Overshoot in Dynamically Switching Inputs. In: Proc. IEEE Intl. Conf. on Emerging Trend in Engineering and Technology, pp. 765–769 (2010)
12. Ndubuisi, E.: Power Dissipation and Interconnect Noise Challenges in Nanometer CMOS Technologies. IEEE Potentials, 26–31 (2010)
13. Shin, Y., Kim, H.O.: Analysis of Power Consumption in VLSI Global Interconnects. In: IEEE International Symp. on Circuits and Systems, vol. 5, pp. 4713–4716 (2005)

14. Kaushik, B.K., Sarkar, S., Agarwal, R.P.: Terminating Load Dependent Width Optimization of Global Inductive VLSI Interconnects. In: IEEE International Conf. on Emerging Technologies, pp. 301–305 (2005)
15. Gargi, K., Chandel, R., Chandel, A.K., Sarkar, S.: Analysis of Non-ideal Effects in Coupled VLSI Interconnects with Active and Passive Load Variations. Microelectronics International 26(1), 3–9 (2009)
16. Shin, Y., Sakurai, T.: Power Distribution Analysis of VLSI Interconnects using Model Order Reduction. IEEE Trans. on Computer-Aided Design 21(6), 739–745 (2002)
17. Delorme, N., Belleville, M., Chilo, J.: Inductance and Capacitance Analytic Formulas for VLSI Interconnects. Electron Lett. 32(11), 996–997 (1996)
18. Lu, Y., Banerjee, K., Celik, M., Dutton, R.W.: A Fast Analytical Technique for Estimating the Bounds of On-Chip Clock Wire Inductance. In: Proc. IEEE Custom Integrated Circuits Conf., pp. 241–244 (2001)

Image Analysis of DETECHIP® – A Molecular Sensing Array

Marcus Lyon[1], Mark V. Wilson[1], Kerry A. Rouhier[2], David J. Symonsbergen[3], Kiran Bastola[4], Ishwor Thapa[4], Andrea E. Holmes[1], Sharmin M. Sikich[1], and Abby Jackson[1]

[1] Doane College, Department of Chemistry, 1014 Boswell Avenue, Crete, NE 68333
abby.jackson@doane.edu, sharmin.sikich@doane.edu
[2] Kenyon College, Department of Chemistry, 200 N. College, Gambier, OH 43022
[3] NOVEL Chemical Solutions, 1155 E. Hwy 33, Crete, NE 68333
[4] University of Nebraska at Omaha, School of Interdisciplinary Informatics, 6001 Dodge Street, Omaha, NE 68182

Abstract. Several image analysis techniques were applied to a colorimetric chemical sensor array called DETECHIP®. Analytes such as illegal and over the counter drugs can be detected and identified by digital image analysis. Jpeg images of DETECHIP® arrays with and without analytes were obtained using a camera and a simple flatbed scanner. Color information was obtained by measuring red-green-blue (RGB) values with image software like GIMP, Photoshop, and ImageJ. Several image analysis methods were evaluated for analysis of both photographs and scanned images of DETECHIP®. We determined that when compared to photographs, scanned images of DETECHIP® gave better results through the elimination of parallax and shading that lead to inconsistent results. Furthermore, results using an ImageJ macro technique showed improved consistency versus the previous method when human eyesight was used as a detection method.

Keywords: DETECHIP, Molecular Sensing Array, Color Signal, Detection of Narcotics, Cutting Agents, Image Analysis, RGB, GIMP.

1 Introduction

Designed for lab and field use, DETECHIP® is a mix-and-measure assay that is capable of providing both colorimetric and fluorescent signals for the rapid detection and identification of molecules of emerging interest such as narcotics, narcotics with cutting agents, over the counter medications, volatile organic compounds, explosives and the intermediates used to make them, microbial metabolites, and environmental contaminants like pesticides [1, 2]. The term DETECHIP® (short for detection chip) combines the concept of small molecule detection with the use of an array of chemical indicators. There are many applications that require a quick, sensitive and selective detection system for specific compounds, including alerting security officers to the presence of explosives and their precursors, screening for weapons of mass destruction, testing biological fluids for illegal compounds, and detection and

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 145–158.
springerlink.com

quantification of sports doping compounds. Colorimetric assays (i.e. "spot tests") offer speed, simplicity of operation, portability, and affordability [3-6]. The stability and versatility of these spot tests enable lab scientists or field personnel to "triage" samples and select those for additional analysis, but they do not provide positive identification.

GC-MS [7-9] is the most widely used method to detect these types of substances, but sample introduction, miniaturization, and the need for skilled operators remain a challenge. Furthermore, high-resolution instruments and expensive additional assays such as isotope ratio mass spectrometry (GC-IRMS) are often required to distinguish between similar compounds. Highly specific tests, such as enzyme-linked immunosorbent assays (ELISA) typically involve chromophore reporters that produce a color, fluorescent, or electro-chemiluminescent change to indicate the presence of specific antigen [10]. While these immunoassays offer extremely high sensitivity, they are also expensive, non-quantitative, and have limited shelf life because they are protein-based and water or humidity sensitive.

None of the described methods are practical for screening thousands of compounds spanning several different molecular classes, and it is this need that DETECHIP® fills. DETECHIP® offers a simple, sensitive, selective, and affordable alternative to existing technologies for the detection of analytes including heroin, cocaine, tetrahydrocannabinoldate (THC) from marijuana, as well as date-rape and club drugs such as flunitrazepam, gamma-hydroxy butyric acid (GHB), or methamphetamine. Significantly, the same system also uniquely identified the explosive trinitrotoluene (TNT), five organic compounds produced by spoilage yeast in beer and wine, as well as over 25 pesticides that are an environmental concern to the U.K. government [11]. Shown to be contactless, portable, inexpensive, DETECHIP® can be adapted to identify a number different classes of substances. Unlike other color tests, which result in a single 'yes' or 'no' response intended to signify that a functional group is present, e.g. the amino group of a narcotic [3, 12], DETECHIP® provides many simultaneous responses, allowing users to quickly characterize and identify suspect materials by assembling a unique, substance specific, binary code composed of '1' and '0'. In this code, '1' represents a change in color or fluorescence, while '0' represents no change. Figure 1 summarizes the assembly and interpretation of DETECHIP®. First, the DETECHIP® sensors, represented by the blue drops in Figure 1a, are placed into the wells of the 96-well plate with each row of twelve containing a unique DETECHIP® sensor. The sensors are then exposed to the analyte of interest, represented by the red drops in Figure 1b. The analyte is added to alternating 8-well columns to provide a control well for later comparison. Figure 1c represents a well that has experienced after analyte addition, and includes a detection method. The ability for the simultaneous detection of controls and suspect materials is unique to DETECHIP®. Figure 1d includes an image of DETECHIP® in its entirety illustrating color changes as well as fluorescent changes when exposed to UV light.

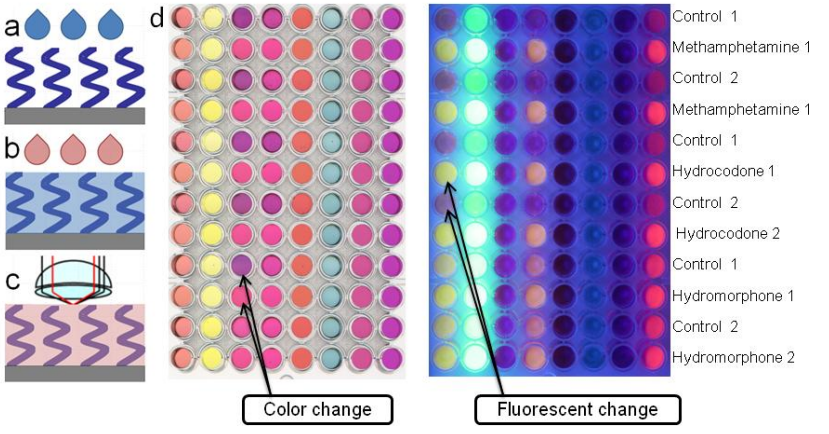


Fig. 1. Illustration of sensing principles for parallel monitoring/readout of molecular interactions on DETECHIP® using image analysis of color images: (a) placement of DETECHIP® elements (blue drops) into 96-well plates (b) exposure of DETECHIP® to analyte (red drops) (c) measurement of RGB change with image software. (d) A typical DETECHIP® ready for analysis. Color (CC) and fluorescence (FC) changes in the sample well relative to the control well are noted (arrows). These changes are recorded as a binary code. A "0" indicates no change while "1" denotes a change in the sample. A representative code for methamphetamine in the presence of an adulterant (baking soda) is 1111-0011-1111-0000-1100-0000-0001.

Affording individual codes for the multitude of compounds listed in Table 1, DETECHIP® has the unique ability to detect and discriminate substances in many difference classes: over-the-counter (OTC) medications, explosives, pesticides, food spoilage metabolites, drugs laced with cutting agents, and various other organic molecules [1-4, 12].

Highly successful in its current form, DETECHIP®'s 32 digit binary code for each analyte is obtained by a person visually inspecting each analyte well and comparing it to the control well. Unfortunately in some tests, differences in human vision and subjective interpretation produce inconsistent codes for identical analytes. This variability can be minimized through plate analysis by multiple people with the consensus determining the binary code. Although providing highly consistent results[13], this method is labor, time, and personnel intensive. In order to circumvent this timely process and eliminate human variability, and thereby decreasing the occurrence of inconsistent codes and false positives, image analysis techniques were employed.

In the last several years, the use of colorimetric sensing, using red-green-blue (RGB) values, as a detection method for digital array images has increased in popularity [13-18]. Various analytes can be detected using RGB color space including pigments of green beans [18], nitrates [16], sugars [14], peroxide vapors [15], and biogenic amines [17]. "Optoelectronic noses" have been reported in conjunction with image analysis and have shown to be an effective detection tool for odorants and gases, but not for abused substances and other analytes of interest to us [19-22]. Used to identify the analyte and determine concentration, these digital image analysis methods are based on color differences as perceived by the software. Examples of software programs that have been used for this type of analysis include ImageJ [23], gimp [24], and Adobe Photoshop.

Table 1. List of substances currently under investigation using DETECHIP®

Drug or Narcotic	Over the Counter	Drugs with Cutting Agents	Pesticides
Phenylcyclohexyl piperidine	24 Hour allergy relief D	Cocaine/Baking soda (1:1)	2,4 - Diiodo-4-hydroxybenzonitrile
Caffeine	24 Hour allergy relief D	Cocaine/Dextrose (1:1)	2-Hydroxy-1-(2-Hydroxy-4-Sulpho-1-Napthlazo)-3-Napthoic Acid
Cocaine	Caffeine	Cocaine/Epsom salt (1:1)	3-(3,4-dichlorophenyl)-1,1-dimethylurea
Codeine	DG Antacid tablet	Cocaine/Glucose (1:1)	3-(4-chlorophenyl)-1-methoxy-1-methylurea
D-Amphetamine sulfate	DHEA (Dehydroepiandrosterone)	Cocaine/Lactose (1:1)	3,5 - Diiodo-4-hydroxybenzonitrile
Fentanyl	Enteric coated aspirin	Cocaine/Lidocane (1:1)	4,7 -Diphenyl-1,10-phenanthroline disulphonic acid disodium salt
Flunitrazepam	Equate allergy medication	Cocaine/Mannitol (1:1)	4-Chloro-o-tolyloxyacetic acid
Hydrocodone	Equate naproxen sodium	Cocaine/Methylsulfone (1:1)	4-Dimethylaminobenzylidene-rhodanine
Hydromorphone	Equate sleep aid	Cocaine/Phenacetin (1:1)	Asulam
Ketamine	Glucosamine Chondroitin	Cocaine/Powdered milk (1:1)	Atrazine
L-Alphacetylmethadol	Ibuprofen	Cocaine/Powdered sugar (1:1)	Bromoxynil
Methadone	Jet-alert	Cocaine/Starch (1:1)	Chlorotoluron
Methamphetamine	L-Glutamine	Cocaine/Sugar (1:1)	DDE
Methylphenidate	Multivitamin	Cocaine/Talc (1:1)	Dichlorprop
Morphine	Phenacetin	Meth/Baking soda (1:1)	Dithizone
Quinine	Suphedrine sinus headache	Meth/Dextrose (1:1)	Diuron
Thebaine	Tylenol cold day	Meth/Epsom salt (1:1)	Endosulfan
	Tylenol cold night	Meth/Glucose (1:1)	Endrin
		Meth/Lactose (1:1)	Gamma-BHC
		Meth/Lidocane (1:1)	Hexamethyldisilazane
		Meth/Mannitol (1:1)	Ioxynil
Explosives	Wine Cork Metabolites	Meth/Methylsulfone (1:1)	Isoproturon
TNT (Trinitrotoluene)	Guaiacol	Meth/Phenacetin (1:1)	Linuron
	Geosmin	Meth/Powdered milk (1:1)	MCPA
	TCA	Meth/Powdered sugar (1:1)	MCPB
	4-Ethylguaiacol	Meth/Starch (1:1)	Mecoprop
	4-Ethylphenol	Meth/Sugar (1:1)	p.p'-DDT
		Meth/Talc (1:1)	Simazine

Utilizing similar photo analysis techniques, the hypothesis for DETECHIP®, is that digital measurements of RGB values are more objective than previous measurements made by visual interpretation and will therefore eliminate errors caused by differences in human vision and subjective interpretation of color. Although the original DETECHIP® format employed both color and fluorescent changes to produce a 32-digit binary code [1, 2, 12], using image analysis will eliminate the fluorescence aspect of DETECHIP®. The code will therefore rely solely on the interpretation of color changes to produce a now 16-digit binary code. This manuscript presents several different digital image analysis techniques for the interpretation of DETECHIP®, using commercially available or in-house designed image software for the measurement of changes in RGB intensities after exposure to the analyte of interest. The resulting codes and their associated accuracy and precision will be compared against codes previously established through visual interpretation. This approach will be advantageous in the progression towards automation of DETECHIP®.

2 Materials

DETECHIP® plates were prepared for triplicate analysis of a single analyte at a concentration of 62.5 mM, as opposed to previous plates that were prepared with three separate analytes. This allowed for the investigation of consistency in the determined codes. Digital images of DETECHIP® were obtained using either a Canon EOS Rebel T1 EOS 500D camera with an EF 50mm f/2.5 compact-macro lens, or an Epson V700 Photo Scanner. Analysis of the resulting digital images was done using ImageJ, and gimp software programs, as well as Adobe Photoshop, MATLAB, Microsoft Excel, and an in-house designed macro.

3 Methods

3.1 Photo Analysis Method

Photos of the DETECHIP® plates were taken using a Canon EOS Rebel T1 EOS 500D camera with an EF 50mm f/2.5 compact-macro lens. Digital photo analysis was performed on a JPEG file that was created using a 10:1 compression algorithm with negligible loss in image quality. The first step of the analysis process included preprocessing of the image file to locate exact regions with in the wells, which contained control and analyte samples. The preprocessing steps included cropping the image file, scaling down the resolution to 500 x 800 pixels, and determining the center of each sample well. A MATLAB program [25] was used to identify the centers of each well. The results, however, were not highly compelling. Therefore, an interpolation program was written in Perl to not only filter out the improperly identified centers but also to detect missing centers. The centers for each of the 96 wells were highly accurate, all being checked visually by plotting the centers determined by the algorithm of the image. A MATLAB program was written to determine the RGB intensities for each pixel in a circular region around these centers and thus the total RGB intensity for every well. The average RGB values were calculated based on four different plates of the same drug/sensor combination. If the error bars (standard deviation) of

the RGB intensities for the analyte did not overlap with those for the controls, it was counted as a statistically significant change and given a '1' for a code, but if there was an overlap of the standard deviation between the analyte and either one or both of the controls, then a '0' was assigned. The codes were then compared between the original visual analysis method, and the new digital analysis protocol. The photo analysis method measures the RGB intensities but not fluorescence, thus in order to make a comparison between the color changes performed by visual inspection with the digital photo analysis, the codes were changed to account only for color changes, thereby reducing the original 32-digit code to the now 16 digit code.

3.2 Scanned Image Analysis

A DETECHIP® 96 well-plate containing three identical tests of one analyte of interest was scanned using an Epson V700 Photo Scanner. The positive film scanned image was 1350x1983 pixels and was saved as a TIFF image. This scanned image was then analyzed using three image analysis techniques, all of which employed the use RGB values. As previously stated, in reference to the photo analysis method, the omission of fluorescence reduced the resulting code from 32 to 16 digits.

3.3 Subtraction Method

For image analysis via the subtraction method, the scanned image, saved as a TIFF file, was opened in gimp (GNU Image Manipulation Program), a free online imaging software program. As seen in Figure 2, two separate images were created for the original scanned TIFF file. In the first, the analyte wells were eliminated, leaving only the control wells, while in the second only the analyte wells remain. These two

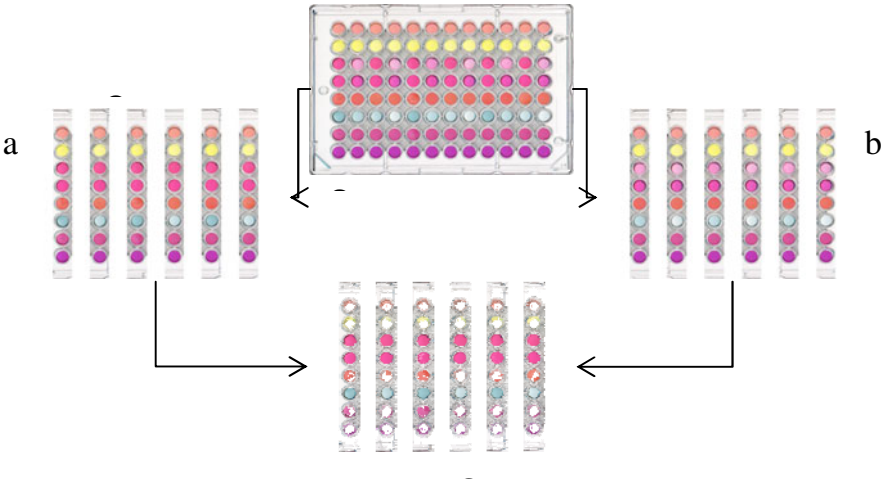


Fig. 2. Example of an image subtraction assay. The scanned image of the control and analyte wells is subtracted from each other. The resulting image subtraction is then analyzed for wells that are colorless versus colored. A colorless well indicates that the control and analyte well are the same and codes for a '0'. A colored well indicates that there is a difference in color between the analyte and color wells and codes for a '1'.

images were then subtracted using the image subtraction filter in the gimp software, setting the analyte image as the master (Figure 2a) and the control image as the slave (Figure 2b). By subtracting out the color of the control well from that of the analyte well, this qualitative method reveals the color difference caused by the addition of analyte. Any color remaining in the well indicates a color change and results in a value of '1' in the final code. If no color change exists, the resulting image will be white, corresponding to a code value of '0'. Interference does exist due to the presence of the wells themselves in the image, but this can be eliminated through the use of image masking. The following method, referred to as the Masking Method, takes the qualitative nature of the image subtraction method and adapts it to produce quantitative method for code determination.

3.4 Masking Method

As seen in Figure 3, the masking method begins with the positive film scanned image of a DETECHIP® 96 well-plate. The gimp software was the used to create a black mask with the elimination of 96 perfect circles, each with a diameter of 68 pixels. The mask is layered overtop the image of the DETECHIP® 96 well-plate. The new masked image was then opened for further analysis in ImageJ. Using the threshold selection tool, all 96 circles were selected and analyzed for average RGB values, area, and standard deviation. Consistency across the three tests was analyzed through comparison of values ± 1 standard deviation. Overlap in these values indicated that the three tests gave statistically identical results. Additionally, a comparison of the control and analyte well for each dye was also done for each of the three tests, altering the standard deviation to facilitate consistency across these three tests. Initially, the three tests were analyzed for consistency at one standard deviation, compiling a code for each test. These three codes were then compared. If discrepancies appeared between the codes, the value of the standard deviation was decreased by a factor of 0.1, after which the values were reanalyzed for overlap and the codes were determined again. This process continues until all three tests resulted in identical codes.

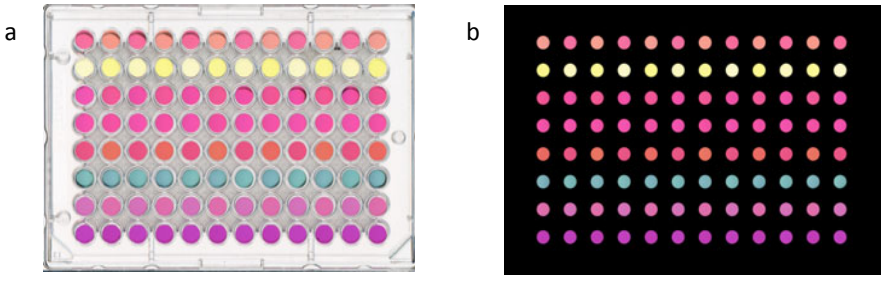


Fig. 3. Scanned images of caffeine on a DETECHIP® plate: (a) original scanned image (b) image after proper orientation and masking.

3.5 Macro Method

After opening a scanned image of a DETECHIP® 96 well-plate in the ImageJ program, the image was properly oriented for analysis using a newly designed macro, created in-house, through modification of previously published work [16]. This macro is designed to select a circular area (47 x 50 pixels) in the center of each well. Within the selected region of the well, each pixel was analyzed to obtain a value for Red. These values were used to calculate an average Red value for the entire selected region of the well. This process was then repeated on the same region to obtain average values for Green and Blue. This sequence of measurements is preformed on all 96 wells. The analysis of both analyte and control wells allowed for simultaneous comparison of these color values. The macro was programmed to produce a '1' if there is a sizeable difference in any of the Red, Green, or Blue values between the control and the analyte wells. The term 'sizable' refers to a value larger than the set color threshold. This threshold value, optimized through experimentation, provides a quantifiable cutoff for what is considered a color change. Differences larger than this threshold are considered sizeable; producing a '1' in the code, while differences smaller than this value characterized as 'no change', producing a '0' in the code.

4 Results

4.1 Photo Analysis Method

Due to the fact that the photo analysis method measures only the RGB color intensities but not fluorescence, the original 32-digit DETECHIP® codes were reduced to 16 digits. This allowed for a direct comparison between the color changes recorded by human visual inspection and the codes determined though digital photo analysis. The average RGB values were calculated based on four different plates of the same drug/molecular sensor combination. Figure 4 shows results of the RGB values of analyte wells versus control wells. If the error bars (plus or minus one standard deviation) of the RGB intensities for the analyte did not overlap with those for the controls, was counted as a statistically significant change and given a '1' for a code. If there was a standard deviation overlap between the analyte and either one or both of the controls, then a '0' was assigned.

Overall, the results of visual inspection versus digital photo analysis matched well when codes produced by the two methods were compared. In some instances, as seen in Table 2, the digital photo analysis indicated a significant color change that was not visible by eye (i.e. methamphetamine at positions 4, 11, and 12 in Table 2).

Highlighted in grey are three cases in contrast, whereby color changes were seen visually, but were not detected by digital photo analysis. As an example, for hydrocodone in buffer B, sensor DC2 (fourth digit in hydrocodone code, Table 2) displays a significant difference in blue intensity between the analyte and the control. For hydromorphone and methadone, the color intensity values were very subtle, while color differences between analyte and control were more difficult to detect.

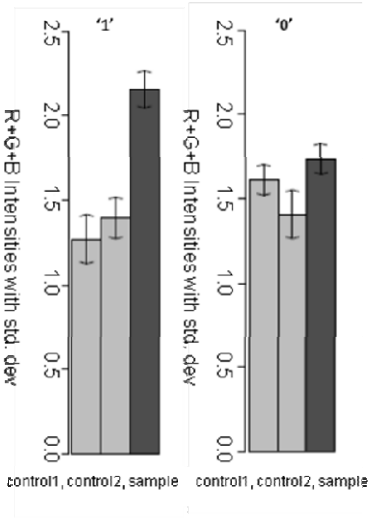


Fig. 4. Example of code assembly using photo analysis method. The error bars on the left are not overlapping, thus the code is "1". The error bars on the right are overlapping and thus code a "0". The resulting codes are compared with codes previously determined using visual interpretation.

Table 2. Analytes from six molecular classes that gave unique 32-digit codes using the 8-sensor Macro-DETECHIP®.

Methamphetamine	
Visual	0100111100000000
Digital	0101111100110000
Hydromorphone	
Visual	1100000000101000
Digital	1100111000110000
Methadone	
Visual	1100111111111101
Digital	1111111111111001
Hydrocodone	
Visual	1111111000110000
Digital	0010111000110000

The photos used for photoanalysis suffered of parallax and shading of the wells which led to large variations and lack of consistency. In order to improve the image quality and consistency of the assay, the plates were scanned with a flatbed scanner in transparency mode, leading to less parallax, more clear and consistent images, less shading differences, and more consistent lighting. Figure 5 demonstrates the improvement of image quality when DETECHIP® was scanned instead of photographed.

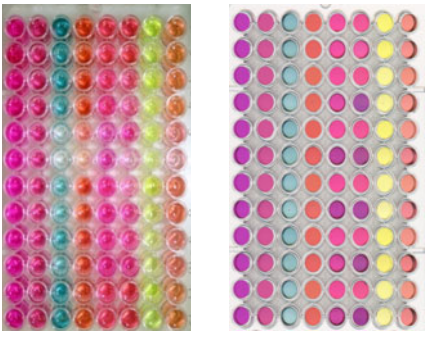


Fig. 5. Left: DETECHIP® photographed with a Canon camera. Right: DETECHIP® scanned with a flatbed scanner.

4.2 Masking Method

This method was built upon the initial findings of the image subtraction method, after it was determined that a more quantitative method was required. The scanned image of the DETECHIP® 96 well-plate was analyzed in ImageJ not only to determine a code for the analyte of interest, but also to look into the consistency between the three tests of the analyte within the plate. For code determination, values for average RGB intensity and standard deviation of RGB intensity was determined for each well. This was accomplished through software analysis of all pixels within each well. The average RGB value and associated standard deviation for each analyte well was compared to the corresponding control using Microsoft Excel. Similar to the photo analysis method, if the error bars – representing ±1 standard deviation value – overlapped, a value of '0' was assigned to the code – signifying that there was no appreciable difference in color between the two wells. However, if the errors bars did not overlap, a value of '1' was assigned to the code, thus designating a color change.

The code for the analyte of interest was determined at several multiples of the standard deviation value, beginning at ±1 standard deviation and reducing this value by a factor of 0.1 for each new code. This was done to obtain consistency between the three identical tests within the 96 well-plate. Reducing the size of the standard deviation shortened the length of the error bars. This affected the overlapping regions of the error between the control and analyte values. When all three tests within the plate produced identical codes, the resulting code and standard deviation were reported as the average code the analyte.

The consistency of the three tests on the plate was also analyzed using a similar technique. Identical wells (e.g., well containing DC1, TRIS buffer and analyte) in each of the three tests were compared using standard deviation values. If the corresponding error bars overlapped, it was determined the tests yielded identical results and were therefore consistent. If the error bars did not overlap, it was determined that the tests were not identical. If this was the result, the suspect test was not used in code determination. Table 3 shows the process of code determination using the Masking Method.

Table 3. Codes determined by color changes (CC) for a plate containing three tests of caffeine. As the standard deviation factor (SDF) was decreased, the three tests converged to a consistent code, with no discrepancies between the three tests.

SDF	Test 1	CC	Test 2	CC	Test 3	CC	Code Differences
1.0	01-11-00-00-00-00-00-	4	01-11-00-00-00-00-00-11	5	00-11-00-00-00-00-00-	4	2
0.9	01-11-00-00-00-00-00-	5	01-11-00-00-00-00-00-11	5	00-11-00-00-00-00-00-	4	1
0.8	11-11-00-00-00-00-00-	6	11-11-00-00-00-00-00-11	6	00-11-00-00-00-00-00-	4	2
0.7	11-11-00-00-00-00-00-	6	11-11-00-00-00-00-00-11	6	10-11-00-00-00-00-00-	5	1
0.6	11-11-00-00-00-00-00-	6	11-11-00-00-00-00-00-11	6	11-11-00-00-00-00-00-	6	0

4.3 Macro Method

This method resulted as an adaptation of the Masking Method. Although thorough and consistent, code determination was quite time intensive. To alleviate this, two method parameters were changed. First, a macro was designed to analyze red, green and blue values separately, as opposed to an average of all three. Second, this macro was also designed to determine the code immediately following RGB measurement. These two improvements alone dramatically decreased analysis time.

Preliminary results using this image analysis technique show improved consistency versus the previous method using human eyesight as a detection method. Studies to determine the reproducibility of this image analysis technique resulted in an average code for each analyte tested. This code was compared against all obtained codes to determine an error percentage. The average codes for three different analytes, along with their associated error percentage, can be seen in Table 4.

Table 4. Average codes and the associated error percentages for three of the tested analytes

ANALYTE	CONCENTRATION	AVERAGE CODE	ERROR
Caffeine	62.5 mM	11-11-11-00-11-00-11-00	7.16%
Caffeine	31.25 mM	11-11-11-00-11-00-11-00	0.21%
Cocaine	62.5 mM	11-11-11-11-11-11-00-00	4.06%
Cocaine	31.25 mM	11-11-11-11-11-11-10-00	1.04%
Nicotine	62.5 mM	11-11-11-11-00-00-00-00	9.75%
Nicotine	31.25 mM	11-11-00-00-00-00-00-00	6.46%

5 Discussion

Originally, visual interpretation of DETECHIP® was used to generate a 32-digit binary code exclusive to a particular analyte. While successful, this method of analysis was not only time, labor, and personnel intensive, but also was also subjective to differences in human vision. Therefore, four new methods of image analysis were developed for more objective code determination. The photo analysis method produced codes that were generally in good agreement with those obtained from visual determination. However, the photos used for this method suffered of parallax and well shading, which had a negative effect on consistency. To overcome this problem, the plates were scanned using a flatbed scanner to create clearer images with less variability between tests.

The scanned images were subject to analysis by the remaining three methods. Image subtraction gave qualitative responses for color change, but some subjective interpretation was still required. If, after subtraction, a well was neither completely white nor completely saturated with color, it must be determined what degree of color change corresponds to a value of '1' in the code. This undesirable quality provided the path to the Masking Method. This method provided a quantitative aspect to image subtraction – providing standard deviation values that could be compared between control and analyte wells. Although consistent codes were obtained, large amounts of analysis time were required.

The Macro Method took aspects of all the previous methods to produce a simple automated analysis technique with incorporated code determination. This method separately analyzed red, green and blue values before assigning a '1' or a '0' value, which was advantageous if one color value increased while a second decreased in a similar fashion. This type of change would is not evident in the average RGB value, and would not be represented in the code determined by the Masking Method. The macro, used within ImageJ, can be modified by setting a color threshold value, altering the size of color change needed to produce a '1' in the code. This quality will be further explored in association with analyte concentration.

Through the analysis of several image collection formats and image analysis techniques, it was determined that scanned images in conjunction with the in-house designed macro for use with ImageJ software provided the most consistent means for DETECHIP® code determination. Furthermore, scanning the DETECHIP® well plates avoids problems associated with photographing the plates, such as parallax and shading around the edges of the wells. These scanned images provide a clear representation of the color in each well, while the macro allows for quantitative determination of color changes between analyte and control wells with consistency that far exceeds the other methods of analysis.

In conclusion, the advantage and value of the DETECHIP® array lies within its ability to detect and identify a multitude of chemicals spanning several different chemical classes, including abused narcotics: narcotics with cutting agents; over the counter medications; explosives and the starting materials or intermediates used to make them; pesticides and other environmental contaminants; metabolites of microorganisms; poisons; etc. We have shown that these analytes of interest can be detected not only by visual inspection of DETECHIP® but also by image analysis, making the detection technique less subjective and more user-friendly. Digital

analysis also opens the door for miniaturization and automation of the DETECHIP® technology.

Acknowledgements. This research was supported in part by the NIH, P20 RR016469 (A.E.H. and M.W.) from the INBRE Programs of the National Center for Research Resources; the NSF CHE-0747949 (A.J.; K.R. and S.S.) and NSF-EPSCoR-EPS-1004094. (A.E.H. and M.L.)

References

[1] Burks, R.M., Pacquette, S.E., Guericke, M.A., Wilson, M.V., Sy-monsbergen, D.J., Lucas, K.A., Holmes, A.E.: Detechipr: A sensor for drugs of abuse. Journal of Forensic Science 55(3), 723–727 (2010)

[2] Holmes, A.E.: Detechip: Molecular color and fluorescent sensory arrays for small molecules. United States Patent US2010/0197516 (2009)

[3] Unidc, P. (ed.): Rapid testing methods of drugs of abuse. United Nations, New York (1994)

[4] O'Neal, C.L., Crouch, D.J., Fatah, A.A.: Validation of twelve chemical spot tests for the detection of drugs of abuse. Forensic Science International 108(1), 189–201 (2000)

[5] Morris, J.A.: Modified bobalt thiocyanate presumptive color test for ketamine hydrochloride. Journal of Forensic Science 52(1), 84–87 (2007)

[6] Justice, U.S.D.O. (ed.): Color test reagents/kits for preliminary identification of drugs of abuse, Washington, D.C (July 2000)

[7] ElSohly, M.A., Salamore, S.J.: Prevalence of drugs used in cases of alleged sexual assault. Journal of Analytical Toxicology 23, 141–146 (1999)

[8] Kollroser, M., Schober, C.: Simultaneous analysis of flunitrazepam and its major metabolites in muman plasma by high performance liquid chromatography tandem mass spectrometry. Journal of Pharmaceutical and Biomedical Assays 28, 1173–1182 (2002)

[9] Huang, Q., He, X., Ma, C., Liu, R., Yu, S., Dayer, C.A., Wenger, G.R., McKernam, R., Cook, J.M.: Pharmacophore/receptor models for gabaa/bzr subtypes ($\alpha1\beta3\gamma2$, $\alpha5\beta3\gamma2$, and $\alpha6\beta3\gamma2$) via a comprehensive ligand-mapping approach. Journal of Medicinal Chemistry 42(1), 71–95 (2000)

[10] Negrusz, A., Moore, C., Deitermann, D., Lewis, D., Kaleciak, K., Kron-strand, R., Feeley, B., Niedbala, R.S.: Highly sensitive micro-plate enzyme immunoassay screening and nci-gc-ms confirmation of flunitrazepam and its major metabolite 7-aminoflunitrazepam in hair. Journal of Analytical Toxicology 23(6), 429–435 (1999)

[11] Sure Screen Diagnostics, Ltd., U.K (2011)

[12] Lyon, M.: Detechipr: An improved molecular sensing array. Journal of Forensic Research 2(4), 1–7 (2011)

[13] Liang, K., Li, W., Ren, H.R., Liu, X.L., Wang, W.J., Yang, R., Han, D.J.: Color measurements for rgb white leds in solid-state lighting using a bdj photodetector. Displays 30(3), 107–113 (2009)

[14] Lim, S.H., Musto, C.J., Park, E., Zhong, W., Suslick, K.S.: A colorimetric sensory array for detection and identification of sugars. Organic Letters 10(20), 4405–4408 (2008)

[15] Lin, H., Suslick, K.S.: A colorimetric sensory array for detection of triacetone triperoxide vapor. Journal of American Chemical Society 132(44), 15519–15521 (2010)

[16] Soldat, D.J., Barak, P., Lepore, B.J.: Microscale colorimetric analysis: Using a desktop scanner and automated digital image analysis. Journal of Chemical Education 86(5), 617–620 (2009)
[17] Steiner, M.-S., Meier, R.J., Duerkop, A., Wolfbeis, O.S.: Chromogenic sensing of biogenic amines using a chameleon probe and the red-green-blue readout of digital camera images. Analytical Chemistry 82(1), 8402–8405 (2010)
[18] Valverde, J., This, H.: Quatitative determination of photosynthetic pigments in green beans using thin-layer chromatography and a flatbed scanner as a densitometer. Journal of Chemical Education 84(1), 1505–1507 (2007)
[19] Feng, L., Musto, C.J., Kemling, J.W., Lim, S.H., Suslick, K.S.: A colorimetric sensor array for identification of toxic gases below permissible exposure limits. Chemical Communications 46(1), 2037–2039 (2010)
[20] Feng, L., Musto, C.J., Suslick, K.S.: A simple and highly sensitive colorimetric detection method for gaseous formaldehyde. Journal of American Chemical Society 132, 4046–4047 (2010)
[21] Janzen, M.C., Ponder, J.B., Bailey, D.P., Ingison, C.K., Suslick, K.S.: Colorimetric sensor arrays for volatile organic compounds. Analytical Chemistry 78(1), 3591–3600 (2006)
[22] Rakow, N.A., Suslick, K.S.: A colorimetric sensor array for odour visualization. Nature 406(1), 710–714 (2000)
[23] Imagej home page (June 2010), http://rsb.info.nih.gov/ij/
[24] Gimp home page (June 2010), http://www.gimp.org/
[25] Peng, T.: Detect circles with various radii in grayscale image via hough transform. MATLAB Central (2005)

A Gaussian Graphical Model Based Approach for Image Inpainting

Krishnakant Verma and Mukesh A. Zaveri

Department of Computer Engineering,
Sardar Vallabhbai National Institute of Technology,
Surat – 395007, Gujarat, India
{krish1931,mazaveri}gmail.com

Abstract. Digital image inpainting means reconstruction of small damaged portions of images. In this paper, we propose an algorithm for digital image inpainting which is a combination of pixel-diffusing technique and a user interaction mechanism. In our approach, the user manually specifies important missing structure information by extending a few curves or line segments from the known region to the unknown regions. Our approach synthesizes image patches along these user-specified curves in the unknown region using patches selected around the curves in the known region. We call this step as structure propagation. After completing structure propagation, we fill in the remaining unknown regions using Gaussian Graphical Model which is MRF based. The experiment results show that our approach is reasonable and efficient. In addition, our method is very simple to be implemented and fast.

Keywords: GGM, MRF, Structure propagation, Image inpainting.

1 Introduction

Image inpainting is a challenging problem in computer graphics and computer vision. Image inpainting aims at filling in missing pixels in a large unknown region of an image in a visual plausible way. Given an input image I with an unknown or missing region Ω, the goal of image inpainting is to propagate structure and texture information from the known or existing regions $I - \Omega$ to Ω, where I is the known region of I. Fig. 1, shows an example of image inpainting technique.

Image inpainting techniques are mainly divided into two categories: pixel-diffusion technique and texture synthesis technique. Pixel diffusion technique is based on partial differential equation (PDE) in image processing and computer vision, which works at pixel level. The basic idea of pixel diffusing is to smoothly propagate information from the surrounding areas in the isophotes direction continuously from exterior [1]. It works well for small region and thin structure, but it has problem with large region because, for large region it does not able to capture texture information and may generate blurring artifacts.

Texture synthesis technique is exemplar based technique which work on patch level. A best matching patch from the known region is patched into unknown region

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 159–166.
springerlink.com

with some automatic guidance. This guidance determines the synthesis ordering, which significantly improves the quality of inpainting by preserving some salient structures [2][3].

Fig. 1. An example of inpainting (a) image with superimposed text. (b) image after inpainting, text is removed.

Actually, the hypothesis for image inpainting is ill-conditioned. Formally, the solving of ill-conditioned problem needs to place constraints on the problem via the Bayesian inference rule. Then the ill-conditioned problem is transformed into a well-conditioned problem. The constraints are generally related to contextual knowledge, which is indispensable for image comprehension and recognition. Additionally, the experiments in neuron-physiology showed that the mechanism of human optical nerves system is closely related to contextual knowledge.

Markov Random Field (MRF) has been used as an appropriate model for the contextual knowledge. Markov random field models provide us "the probabilistic image processing scheme" which may have robustness more than some deterministic approaches with digital filters. In MRF model, the state of a pixel depends only on the configuration of its nearest neighbor pixels [4]. In paper [5], an approach is presented for image restoration to still image based on Gaussian Markov Random Field (GMRF). Further this work is extended by [6] to multi scale GMRF in which local information is covered by GMRF and global information is covered from multi scale images. Other work for image inpainting based on MRF is discussed in [7] under the framework of Bayesian inference. But this algorithm still has some problem for texture and structure information. This algorithm is not able to recover the edge after removing occluded object and to preserve texture information for large region. In [8] authors come up with interactive approach in which user draw the curve from known region to unknown region and then structure propagation to synthesize regions with salient structures. Finally, the optimization problem is solved by Belief Propagation. But it was computationally costlier because the belief propagation algorithm can treat any discrete gray levels image processing, it needs a large amount of computation time in high gray levels case.

In our system, the user manually specifies important missing structure information by drawing some curves from the known to the unknown regions. Our approach

synthesizes image structure along these user specified curves in the unknown region using structure information selected around the curves in the known region. Built on the inpainted structure, we fill in the remaining unknown regions using the Gaussian Graphical Model, instead of belief propagation as used by [8]. This paper is organized as follows: In Section 2, the MRF inpainting algorithm is introduced. Our approach based on GGM inpainting algorithm is presented in Section 3. Section 4 demonstrates the experimental results and analysis. Finally, conclusion is given in Section 5.

2 MRF Image Inpainting Algorithm

2.1 MRF Inpainting Model

The idea of MRF inpainting algorithm is shown in fig. 2. The shadow (yellow) grids represent the unknown regions and the white grids are the known regions. The neighborhood system is defined as a 4-neighborhood one for computational efficiency. The MRF inpainting algorithm works from the pixels of outer boundary to the pixels of inner region.

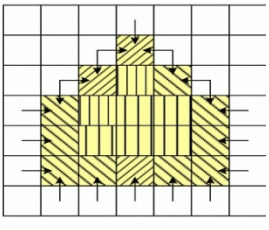


Fig. 2. Pixel propagation

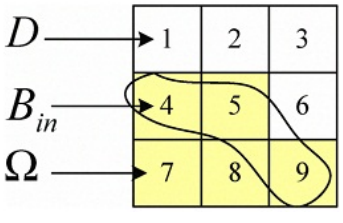


Fig. 3. MRF inpainting Model

Fig. 3 shows the MRF model described in [7] where the shadow (yellow) grids (including 4, 5, 7, 8, 9 pixels) are unknown from the original image and the white grids are known (including 1, 2, 3, 6 pixels). The unknown region to be inpainted is denoted as Ω, and the known region is denoted as D. B_{in} ($\in \Omega$) is a region which is a border of Ω. In order to inpaint the unknown region, gray levels in unknown region need to be estimated based on gray levels of the known region. The energy function based on the MRF model is defined as follows:

$$E(\boldsymbol{f}) = \frac{1}{2}\sum\nolimits_{j\in B_{in}} \beta_i (f_i - z_i)^2 + \frac{1}{2}\sum\nolimits_{ij\in \mathrm{N}\Omega} \alpha_{ij}(f_i - f_j)^2 \quad (1)$$

where $f_i \in \{0, 1, \cdots, Q\}$ denotes a gray level of pixel i($\in\Omega$) and $f = \{f_i \mid i \in \Omega\}$. N Ω is a set of the nearest neighbor pairs of pixels in region Ω. Both parameters $\{\beta i\}$ and $\{\alpha ij\}$ are positive and $\alpha_{ij} = \alpha_{ji}$. z_i is defined as follows:

$$z_i = \frac{1}{|\mathrm{N}_{D(i)}|}\sum\nolimits_{j\in \mathrm{N}(i)} g_j \quad (2)$$

where g_j is a gray level at pixel j which is in region D and, $\mathrm{N}_D(i)$ is a set of pixels which are nearest neighbor of pixel i($\in B_{in}$) and are in region D. z_i expresses the mean

value of gray level of pixels, which neighbor pixel i($\in B_{in}$), in region D. Since $\{gi\}$ is given from an image, $\{zi\}$ is also given as a data. The first term of (1) expresses the relationship between inside and outside of a border of region Ω and the second term expresses the relationship between nearest neighbor pixels in region Ω. Parameters $\{\beta i\}$ and $\{\alpha ij\}$ adjust them.

To apply our model to the framework of the probabilistic image processing, we introduce a probability distribution:

$$\mathrm{P(f)} = \frac{1}{\mathrm{Z}}\mathrm{e}^{-\mathrm{E(f)}} \tag{3}$$

where Z is a normalization constant. The probability distribution is referred to as a Gibbs distribution and it is hard to treat it exactly. In this framework, the goal can be achieved by

$$\dot{\mathrm{f}} = \arg\ max_f P(f) \tag{4}$$

where $\dot{\mathrm{f}}$ are estimated values of gray levels of the inpainted region Ω. The maximum of probability distribution (4) corresponds to the minimum of the cost function (1).

2.2 Gaussian Graphical Model for Inpainting

In practical digital images, each pixel usually has 256 gray levels, i.e., 0, 1, 2, ⋯ , 255. We assume such gray levels can be regarded as continuous values approximately. Actually, this assumption is adopted in many conventional digital image filters [9]. Under this assumption, our presented model becomes Gaussian graphical model (GGM) which can be treated exactly [10]. In this section, we assume each $\mathrm{f_i}$ takes any real number in the interval (–∞, ∞), therefore, we can solve this problem analytically. The probability distribution (3) is expressed as the following form:

$$\mathrm{P(f)} = (2\pi)^{-\frac{|\Omega|}{2}}(\mathrm{detA})^{\frac{1}{2}} * \exp\{-\frac{1}{2}(\mathrm{f}-\mathrm{m})^{\mathrm{T}}\mathrm{A}(\mathrm{f}-\mathrm{m})\} \tag{5}$$

where matrix $\mathrm{A} = \{\mathrm{A_{ij}} \mid \mathrm{i} \in \Omega, \mathrm{j} \in \Omega\}$ and vector $\mathrm{m} = \{ m_{\mathrm{i}} \mid {}_i \in \Omega\}$ are defined as follows:

$$\mathrm{A_{ij}} = \begin{cases} -\alpha_{\mathrm{ij}} & \left((\mathrm{ij}) \epsilon \mathrm{N}\ \Omega\right) \\ \dot{\mathrm{B}}_{\mathrm{i}} + \sum_{\mathrm{k}\epsilon \mathrm{N}\ \Omega(\mathrm{i})} \alpha_{\mathrm{ik}} & (\mathrm{i} == \mathrm{j}) \\ 0 & (\mathrm{otherwise}) \end{cases} \tag{6}$$

$$\mathrm{m} = \mathrm{A}^{-1}\mathrm{b} \tag{7}$$

where $\mathrm{b} = \{b_{\mathrm{i}} \mid i \in \Omega\}$ and

$$\dot{\mathrm{B}}_{\mathrm{i}} = \begin{cases} \beta_{\mathrm{i}}, & (\mathrm{i} \epsilon \mathrm{B_{in}}) \\ 0 & (\mathrm{i} \epsilon \mathrm{B_{in}}) \end{cases} \tag{8}$$

$$b_i = {}_i.\mathrm{z}_i \tag{9}$$

P(*f*) takes a maximal value, when f = m. Therefore, our goal (4) is as follows

$$m = \arg\ \max_f P(f) \tag{10}$$

hence the estimated value of gray level of pixel i in Ω is given by $\dot{f}_i=[m_i + 0.5]$, where the notation [x] is an integer less than x. Note that the maximal value of P(f) can be determined uniquely, since the probability distribution (5) is a convex function on f space.

3 Our Approach Based on GGM Algorithm

Since the image processing based on MRF is a probabilistic processing, it potentially could deal with uncertainties in the inpainting problem and lead to a more robust result than deterministic approaches. But some problems still exist for the MRF inpainting algorithm. Usually the image processing techniques under the pixel-wise framework could likely lead to a result where the processed pixel takes the same gray level as surrounding pixels. Thus, images likely become smooth by this kind of processing. In the smoothing region, this assumption may be valid. However, in the region which contains explicit structure information, for instance edges, this assumption is not always valid.

As shown in fig. 4, result is generated by GGM inpainting algorithm [7], we notice that the damaged region which contain edge or structure information is not inpainted well it get blurred by surrounding structure. In Fig. 5 we can see that the damaged region, which contains no structure information or no edge to reconstruct, is well inpainted. The reason for such a result is due to the fact that the MRF inpainting algorithm is a local algorithm working pixel-wisely. It only passes the message along the local MRF chain and does not take the global structure information into consideration. Thus the blurring result in Fig. 4 (b) is unavoidable.

(a)  (b) 

Fig. 4. Result from GGM algorithm (a) the region indicated black is to be inpainted, in which edge needs to be reconstructed. (b) The edge between sky and sea is not preserved.

To take the global information into consideration in the inpainting algorithm, it is essential to include the structure information. Human being is superior and efficient in identifying structure information in unknown regions. Thus including the high level human knowledge into the inpainting algorithm can combine the local geometrical and global structure information together.

In our approach, user draw a line through simple curve drawing interface for much complex and non repeated structure because user knowledge can make the connectedness of structure simple and exact. Fig. 5 shows the approach of our algorithm. It includes following steps:

- First, the curve is drawn from known region to unknown region.
- Second, the patch contain the structure information from the known region is propagated into the missing region along the curve drawn in missing region.
- Finally, the remaining missing region is inpainted by GGM algorithm.

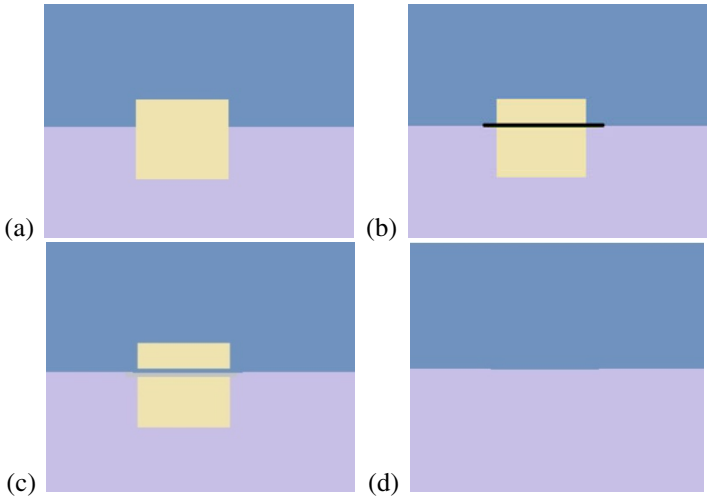


Fig. 5. It shows the process of our GGM based approach. (a) Original image contain missing region. (b) Human draws a line (black) from known to unknown region. (c) Result from structure propagation. (d) Final result after inpainting the remaining missing region by GGM.

Compared with the GGM inpainting algorithm presented in section II, our GGM based approach for inpainting, largely reduces the breaking of salient structures which human eyes are sensitive to. It should be noticed that although the user interaction is the same as in the inpainting algorithm proposed in [2], it is different in filling the remaining unknown regions as in [8], [2]. In this way, our algorithm is a pixel-based while the algorithm in [2] is textual based.

4 Experimental Result

In our experiment, 4-neighbourhood system is considered for GGM algorithm. The weight coefficient is selected manually $\beta_i = 1$ and $\alpha_{ij} = 0.0025$. We applied our algorithm on R, G and B space of image independently. All experiments have carried out on 2.1GHz PC.

We made a comparison between GGM algorithm and our GGM based approach. In Fig. 6, two columns of RGB images of size 256 x 185 each. In right side image

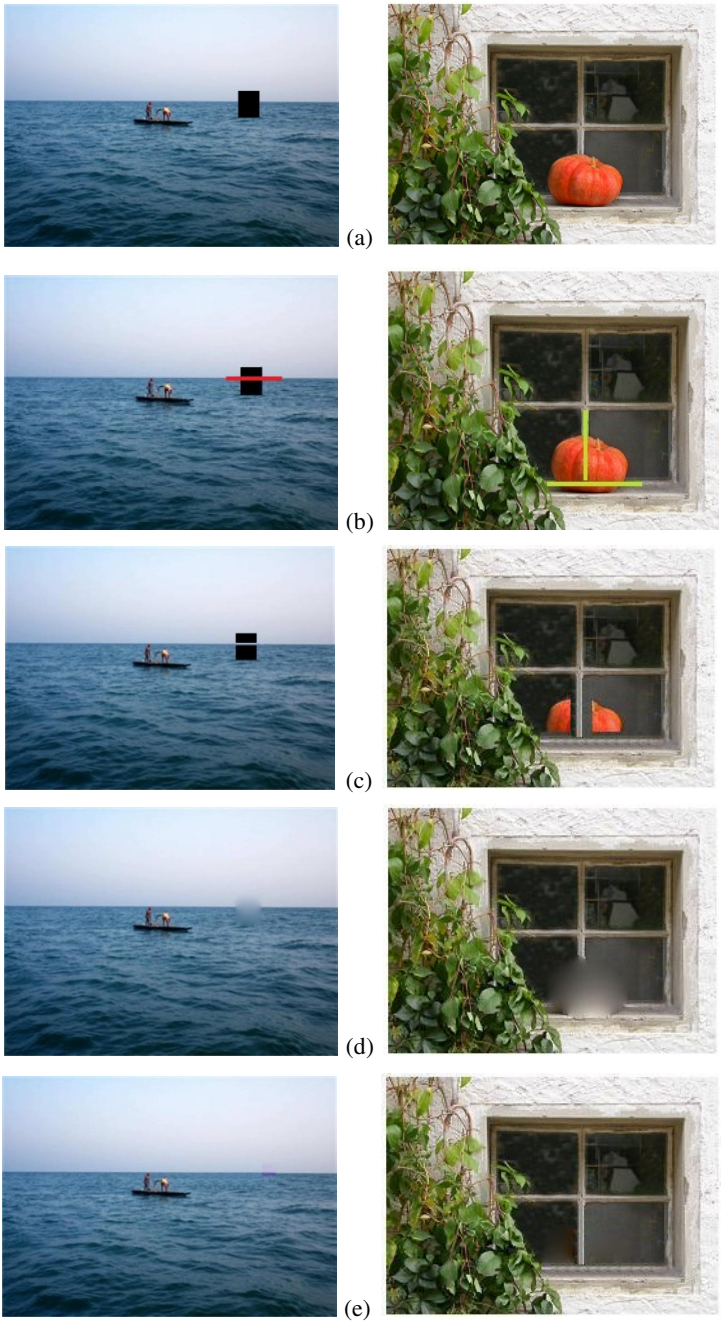


Fig. 6. Performance comparison of GGM algorithm [7] and our GGM based approach. (a) Left image black is unknown region, right image pumpkin has to be removed. (b) Human draw line manually red and green respectively. (c) Result after structure propagation. (d) Final result from our algorithm. (e) Result from GGM algorithm.

pumpkin has to be removed and the task of inpainting is to correctly reconstruct the edge behind the pumpkin. Similarly, in left side image black rectangle has to be removed and inpainting task is to reconstruct the edge between sea and sky.

The results inpainted by our approach are shown in fig. 6(d). The most salient structures are seamlessly propagated from the known region into the unknown region. Inpainted structures look natural. The final results are visually pleasing. However, the results inpainted by the GGM algorithm [7], which are shown in fig. 6(e) contains obvious blurring regions. This is due to the fact that the GGM inpainting algorithm does not take the structure information into consideration and only local pixel neighborhood information is utilized.

5 Conclusion

In this paper, we have presented a new approach based on GGM inpainting algorithm. This is an improvement of the existing GGM [7] inpainting algorithm. Through a curve-based interface, the user indicates which important structures should be completed before remaining unknown regions are filled in. Then, inpainting process is used to produce good results.

References

1. Marcelo, B., Guillermo, S., Vicent, C., Coloma, B.: Image Inpainting. In: Proceeding of the ACM SIGGRAPH 2000, New Orleans, pp. 417–424 (2000)
2. Jian, S., Lu, Y., Jiaya, J., Heung, Y.S.: Image completion with structure propagation. ACM Trans. Graph. 24(3), 861–868 (2005)
3. Antonio, C., Patrick, P., Kentaro, T.: Region filling and object removal by exemplar-based image inpainting. IEEE Transaction on Image Processing 13(9), 1200–1212 (2004)
4. Haluk, D., Howard, E., Roberto, C., Donald, G.: Bayes smoothing algorithms for segmentation of binary images modeled by Markov random fields. IEEE Trans. Pattern Anal. Mach. Intell. 6, 707–720 (1984)
5. Takahiro, O., Miki, H., Hideo, K.: Restoration Method of Missing Areas in Still Images Using GMRF Model. In: International Symposium on IEEE, pp. 931–4934 (2005)
6. Rui, W., Dongbing, G., Guangwen, L., Junxi, S.: A multi-scale image inpainting algorithm based on GMRF model. In: IEEE International Conference on Robotics and Biomimetics (ROBIO), December 19-23, pp. 1844–1848 (2009)
7. Yasuda, M., Ohkubo, J., Tanaka, K.: Digital Image Inpainting based on Markov Random Field. In: International Conference on Intelligent Agents, Web Technologies and Internet Commerce and International Conference on Computational Intelligence for Modelling, Control and Automation, November 28-30, vol. 2, pp. 747–752 (2005)
8. Junxi, S., Defang, H., Dongbing, G., Guangwen, L., Hua, C.: An interactive image inpainting algorithm based on Markov Random Field. In: International Conference on Mechatronics and Automation, ICMA 2009, August 9-12, pp. 101–106 (2009)
9. Jae, S.L.: Two-Dimensional Signal and Image Processing. Prentice Hall PTR (1990)
10. Kazuyuki, T.: Statistical-mechanical approach to image processing. J. Phys. A: Math. Gen. 35, 81–150 (2002)

A Survey on MRI Brain Segmentation

M.C. Jobin Christ[1] and R.M.S. Parvathi[2]

[1] Adhiyamaan College of Engineering, Dr. MGR Nagar, Hosur, India
jobinchrist@gmail.com
[2] Sengunthar College of Engineering, Tiruchengode, India
rmsparvathi@india.com

Abstract. With the growing research on medical image segmentation, it is essential to categorize the research outcomes and provide readers with an overview of the existing segmentation techniques in medical images. In this paper, different image segmentation techniques applied on magnetic resonance brain images are reviewed. The selection of papers includes sources from image processing journals, conferences, books, dissertations and thesis. The conceptual details of the algorithms are explained and mathematical details are avoided for simplicity. Both broad and detailed categorizations of reviewed segmentation techniques are provided. The state of art research is provided with emphasis on developed algorithms and image properties used by them. The methods defined are not always mutually independent. Hence, their inter relationships are also stated. Finally, conclusions are drawn summarizing commonly used techniques and their complexities in application.

Keywords: MRI, Segmentation, Medical Imaging.

1 Introduction

Image processing consists of various application fields like compression, enhancement, detection, feature extraction, restoration, scaling, segmentation, etc. Image segmentation is used in various applications like Medical imaging, locating objects in satellite images, face recognition, traffic control systems, fingerprint recognition and machine vision etc. Medical imaging includes locating tumors and other pathologies, measuring tissue volumes, etc. Segmentation plays an important role in biomedical image processing. Segmentation is the starting point for other processes such as registration, shape analysis, visualization and quantitative analysis. Segmentation of an image is the division or separation of the image into disjoint regions of similar attribute. In clinical practice, Magnetic Resonance Imaging is used to distinguish pathologic tissue from normal tissue, especially for brain related disorders. Three main regions of brain, White Matter (WM), Gray Matter (GM) and Cerebrospinal Fluid (CSF) are the important subject of study in brain imaging. Manual segmentation by an expert is time consuming and it is very difficult to do accurate segmentation. Hence automatic segmentation algorithms are preferred in diagnostic process.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 167–177.
springerlink.com

2 Image Segmentation Methods

Image segmentation algorithms are classified into two types, supervised and unsupervised. Unsupervised algorithms are fully automatic and partition the regions in feature space with high density. The different unsupervised algorithms are Feature-Space Based Techniques, Clustering (K-means algorithm, C-means algorithm, E-means algorithm), Histogram thresholding, Image-Domain or Region Based Techniques (Split-and-merge techniques, Region growing techniques, Neural-network based techniques, Edge Detection Technique), Fuzzy Techniques, etc. It is essential to know which method is to be applicable for the segmentation of medical images. In this paper we present a comparative study of unsupervised algorithms in terms of robustness, accuracy [3, 5, 8].

2.1 Clustering Methods

Clustering is a method of grouping a set of patterns into a number of clusters such that similar patterns are assigned to one cluster. Each pattern can be represented by a vector having many attributes. Clustering technique is based on the computation of a measure of similarity or distance between the respective patterns. In this paper we are going to discuss about K-means algorithm, Fuzzy C-means algorithm.

2.1.1 K-Means Algorithm

K-means algorithm is under the category of Squared Error-Based Clustering (Vector Quantization) and it is also under the category of crisp clustering or hard clustering. K-means algorithm is very simple and can be easily implemented in solving many practical problems. Steps of the K-means algorithm are given below.

1. Choose k cluster centers to coincide with k randomly chosen patterns inside the hyper volume containing the pattern set.(C)
2. Assign each pattern to the closest cluster center.(C_i, i = 1,2,. . . . C)
3. Recompute the cluster centers using the current cluster memberships.(U)
4. If a convergence criterion is not met, go to step 2 with new cluster centers by the following equation, i.e., minimal decrease in squared error.

$$\begin{cases} 1 \text{ if } \|x_j - C_i\|^2 \le \|x_j - C_k\|^2, \text{ for each } k \neq i \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

$$C_i = \frac{1}{|G_i|} \sum_{k,\, x_k \in G_i} x_k \tag{2}$$

Where, $|G_i|$ is the size of G_i

or,

$$|G_i| = \sum_{j=1}^{n} U_{ij} \tag{3}$$

The performance of the K-means algorithm depends on the initial positions of the cluster centers. This is an inherently iterative algorithm. And also there is no guarantee about the convergence towards an optimum solution. The convergence centroids vary with different initial points. It is also sensitive to noise and outliers. It is only based on numerical variables [1, 4, 6, 9].

2.1.2 Fuzzy C-Means Algorithm

Fuzzy C-Means clustering (FCM), also called as ISODATA, is a data clustering method in which each data point belongs to a cluster to a degree specified by a membership value. FCM is used in many applications like pattern recognition, classification, image segmentation, etc. FCM divides a collection of n vectors c fuzzy groups, and finds a cluster center in each group such that a cost function of dissimilarity measure is minimized. FCM uses fuzzy partitioning such that a given data point can belong to several groups with the degree of belongingness specified by membership values between 0 and 1. This algorithm is simply an iterated procedure. The algorithm is given below.

1) Initialize the membership matrix **U** with random values between 0 and 1.
2) Calculates c fuzzy cluster center c_i, $i = 1,\ldots c.$, using the following equation,

$$c_i = \frac{\sum_{j=1}^{n} u_{ij}^{m} x_j}{\sum_{j=1}^{n} u_{ij}^{m}} \tag{4}$$

3) Compute the cost by the following equation. Stop if either it is below a certain threshold value or its improvement over previous iteration.

$$J(U, c_1, \ldots, c_c) = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \sum_{j}^{n} u_{ij}^{m} d_{ij}^{2} \tag{5}$$

4) Compute a new **U** by the equation. Go to step 2.

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{kj}} \right)^{2/(m-1)}} \tag{6}$$

Like K-means clustering, there is no guarantee ensures that FCM converges to an optimum solution. The performance is based on the initial cluster centers. FCM also suffers from the presence of outliers and noise and it is difficult to identify the initial partitions [1, 6, 9].

2.2 Edge Detection Methods

Edge detection is a method which is extensively used for gray level image segmentation. It is a process of finding the discontinuities of an image. Edge detection

is under the category of Boundary based technique. Boundary based methods find connected regions based on finding pixel differences of the pixels within them. The objective is to find a closed boundary such that an outside can be determined easily. Edge Detection process is classified into two broad groups; (i) Derivative approach, (ii) Pattern fitting approach. Both approaches have advantages as well as disadvantages. And also second method gives better result as compared with the first method. Pattern fitting approach uses a series of edge approximation functions over a small neighborhood and it will be analyzed. In derivative approach, edge pixels are found by taking derivatives. Here edge masks are used to find two dimensional derivatives. In this chapter we will discuss about the Roberts, Prewitt, Sobel, Canny, Gaussian, LoG operators in detail. They are under the category of Derivative approach [3, 4, 8].

2.2.1 Roberts Operator

This method is based on differences between adjacent pixels. Here, +1 and -1 are explicitly used to find the edges. This difference is called as forward differences. The first order partial derivative is implemented by cross-gradient operator.

$$d_1 = g_0 - g_2 \tag{7}$$

$$d_2 = g_1 - g_3 \tag{8}$$

The above two partial derivatives are implemented by approximating them into 2x2 windows. The Roberts masks are,

0	1
-1	0

-1	0
0	1

Roberts windows

2.2.2 Prewitt Operator

In Prewitt operator, similar weights are assigned to all the neighbours of the chosen pixel. The first order derivative is given by,

$$d_1 = 1/3[(g_4 + g_5 + g_6) - (g_2 + g_1 + g_8)] \tag{9}$$

$$d_2 = 1/3[(g_8 + g_7 + g_6) - (g_2 + g_3 + g_4)] \tag{10}$$

Its masks are given below,

-1	-2	-1
0	0	0
1	2	1

-1	0	1
-2	0	2
-1	0	1

Prewitt windows

2.2.3 Sobel Operator

In Sobel operator, higher weights are assigned to the close neighbours of the chosen pixel. The first order derivative is given by,

$$d_1 = 1/4[(g_4 + 2g_5 + g_6) - (g_2 + 2g_1 + g_8)] \quad (11)$$

$$d_2 = 1/4[(g_8 + 2g_7 + g_6) - (g_2 + 2g_3 + g_4)] \quad (12)$$

Sobel Masks are given below,

-1	0	1
-1	0	1
-1	0	1

-1	-1	-1
0	0	0
1	1	1

Sobel windows

2.2.4 Canny Operator

It is the most popular operator among all the edge detection algorithms. Canny algorithm mainly concentrate on three things, Maximizing Signal to Noise Ratio (SNR), localization of edges by minimizing the variance of the zero crossing position, identification of single edge rather than multiple response. The canny algorithm is given below,

i. Apply derivative of Gaussian
ii. Non-maximum suppression
iii. Linking and thresholding

2.2.5 LoG Operator

Laplacian of Gaussian operator is otherwise called as Marr-Hildreth operator. It is based on the second derivative method for the detection of zero crossing method. Here in addition to the Laplacian operator, Gaussian smoothing is applied. Laplacian mask is given below,

0	1	0
1	-4	1
0	1	0

Laplacian Window

LoG algorithm is given below.

i. Smooth the image by convolving it with a digital mask.
ii. Apply the Laplacian mask.
iii. Find the zero crossings by Laplacian second derivative operator.

2.3 Watershed Method

Watersheds are one of the typical regions in the field of topography. A drop of the water falling it flows down until it reaches the bottom of the region. Monochrome image is considered to be an height surface in which high-altitude pixels correspond to ridges and low-altitude pixels correspond to valleys. This suggestion says if we have a minima point, by falling water, region and the frontier can be achieved. Watershed uses image gradient to initial point and region can get by region growing. The accretion of water in the neighborhood of local minima is called a catchment basin. Watershed refers to a ridge that divides areas shattered by different river systems. A catchment basin is the environmental area draining into a river or reservoir. If we consider that bright areas are high and dark areas are low, then it might look like the plane. With planes, it is natural to think in terms of catchment basins and watershed lines. Two approaches are there to find watershed of an image,

1. Rainfall approach
2. Flooding approach

In rainfall approach, local minima are found all through the image, and each local minima is assigned an exclusive tag. A intangible water drop is placed at each untagged pixel. The drop moves to low amplitude neighbor until it reaches a tagged pixel and it assumes tag value. In flooding approach, intangible pixel holes are pierced at each local minima. The water enters the holes and takings to fill each catchment basin. If the basin is about to overflow, a dam is built on its neighboring ridge line to the height of high altitude ridge point. These dam borders correspond to the watershed lines. The following steps are used in Watershed Algorithm:

i. Read an Image and covert it into grayscale
ii. Use gradient magnitude as the segmentation function
iii. Mark the foreground objects
iv. Calculate the Background markers
v. Calculate the watershed transform of the segmentation function
vi. Visualize the result

The main drawback of this algorithm is over segmentation, because all of edge and noise would appear in the image gradient. If the signal to noise ratio is not high enough at the contour of interest, the transform won't detect it correctly. It also failed to detect thin structures [4, 5, 8].

2.4 Region Based Methods

Regions are group of connected pixel elements with similar properties. In this method each pixel element is assigned to a particular region. Region growing is a process that groups pixels or sub regions into larger regions. In which nearest pixel elements are examined and added to a region if no edges are detected. It starts with a set of "seed" points and from these produces regions by adding to each seed points those nearest

pixels that have similar properties. Region splitting is another region based approach. It starts with a entire image and divides it into homogeneous regions. Splitting method alone not sufficient for segmentation process. Therefore merging will be applied after splitting, which is called as split and merge method. Steps of Split and merge algorithm is given below.

1. If the entire region is consistent, leave it unchanged.
2. If the region is not sufficiently consistent, split it into four quadrants.
3. Merge any adjacent regions that are similar enough.
4. Repeat steps 2 and 3 repeatedly until no more splitting or merging arises [7, 8].

2.5 Thresholding Methods

Thresholding methods give segments having picture elements with similar gray levels. This technique requires that an object has homogenous gray level and a background with a different gray level. That kind of image can be segmented by two regions using thresholding. Thresholding techniques are classified into Global or fixed thresholding, adaptive thresholding and histogram based thresholding. In this chapter we described OTSU method which is under the category of histogram based thresholding. This method is simple and is an outstanding method for selecting the threshold. For a gray scale image, the total number of pixels is defined as N, n_i is the number of pixels which's intensity is i. By regularizing the histogram, the following equations could be attained.

$$\sum_{i=0}^{255} n_i = N \tag{13}$$

$$p_i = \frac{n_i}{N} \tag{14}$$

p_i is the probability of the pixels which's intensity is i . The threshold of the image segmentation is defined as m, then the probability θ_o and mean value μ_o of the background can be attained through the following equations:

$$\theta_o = \sum_{i=0}^{m} p_i \tag{15}$$

$$\mu_o = \frac{\sum_{i=0}^{m} i p_i}{\theta_o} \tag{16}$$

probability and typical value of the target also can be obtained:

$$\theta_1 = \sum_{i=m+1}^{255} i p_i \tag{17}$$

$$\mu_1 = \frac{\sum_{i=m+1}^{255} ip_i}{\theta_1} \tag{18}$$

By computing all the above values, the following equation is attained,

$$\sigma_B^2 = \theta_0\theta_1(\mu_0 - \mu_1)^2 \tag{19}$$

The threshold which makes the variance yields maximal is the optimal threshold [4, 5].

2.6 Other Methods

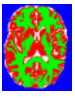
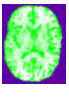
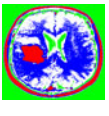
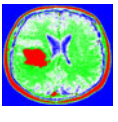
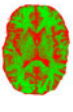
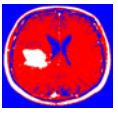
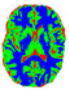
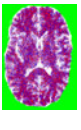
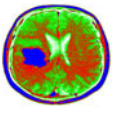
In addition to the above mentioned algorithms, Texture based methods, Wavelet based methods, Level set methods, Wavelength based method, Genetic algorithm based method, neural network based methods, etc. also used for medical image segmentation. Each method is having its own advantages as well as limitations [2].

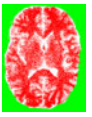
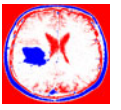
3 Execution

The above methods are implemented using MATLAB 7.9.0(R2009b). MRI brain images are taken for implementation. JPEG and PNG image file formats are used.

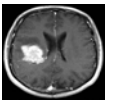
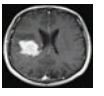
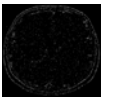
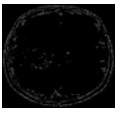
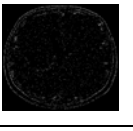
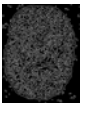
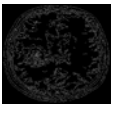
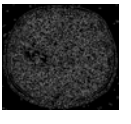
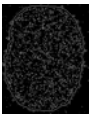
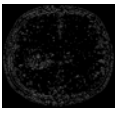
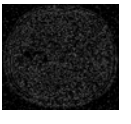
4 Results

4.1 Results of Clustering Methods

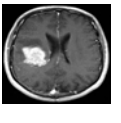
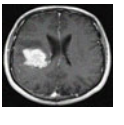
Input Image	Without noise	With Gaussian noise	Without noise	With Gaussian noise
FCM with 4 cluster centres				
FCM with 3 cluster centres				
KM with 4 cluster centres				

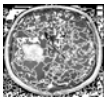
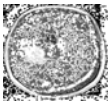
KM with 3 cluster centres				

4.2 Results of Edge Detection Methods

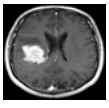
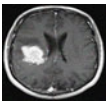
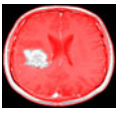
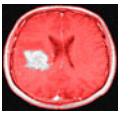
Input Image	 **Without noise**	 **With Gaussian noise**	 **Without noise**	 **With Gaussian noise**
Roberts Operator				
Prewitt Operator				
Sobel Operator				
Canny Operator				
LoG Operator				

4.3 Results of Watershed Method

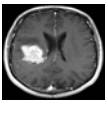
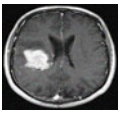
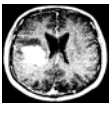
Input Image	 **Without noise**	 **With Gaussian noise**	 **Without noise**	 **With Gaussian noise**

Output Image				

4.4 Results of Split and Merge Method

Input Image	 **Without noise**	 **With Gaussian noise**	 **Without noise**	 **With Gaussian noise**
Output Image				

4.5 Results of Thresholding Methods

Input Image	 **Without noise**	 **With Gaussian noise**	 **Without noise**	 **With Gaussian noise**
Output Image				

5 Conclusions

The different segmentation methods have been demonstrated. From the results, the clustering algorithms are guaranteed to converge but it may not return optimal solution. In K-Means algorithm the quality of the solution depends on the initial set of clusters and value of K. An inapt choice of K yields very poor result. But for White matter segmentation, it gave better results. In noisy environment FCM gave better results than KM. From the results of edge detection techniques, the canny operator performed well than all other operators. But in noisy environment it failed to converge.In watershed algorithms, the length of gradients is taken as elevation information. The flooding process is performed over gradient image, this leads to an over segmentation of an image, especially for noisy environment. Split and Merge

technique operated well over all images. This method performs well even in noisy environment. Mainly it is suitable for detection of tumors, etc. Thresholding technique is not suited for WM, GM and CSF Segmentation. But it is giving satisfactory results for tumor affected images.

References

1. Jain, A.K., et al.: Data Clustering: A Review. ACM Computing Surveys 31(3) (1999)
2. Chaudhuri, B.B., Sarkar, N.: Texture Segmentation Using Fractal Dimension. IEEE Transactions on Pattern Analysis and Machine Intelligence 17(1) (1995)
3. Chanda, B., Dutta Majumder, D.: Digital Image Processing and Analysis. Prentice Hall of India Pvt. Ltd. (2008)
4. Prasantha, H.S., et al.: Medical Image Segmentation. International Journal on Computer Science and Engineering 02(04) (2010)
5. Gonzalez, R.C., Woods, R.E.: Digital Image processing, 2nd edn., pp. 589–656. Pearson Education (2007)
6. Xu, R., Wunsch II, D.: Survey of clustering algorithm. IEEE Transactions on Neural Networks 16(3) (2005)
7. Hojjatoleslami, S.A., Kittler, J.: Region Growing: A New Approach. IEEE Transactions on Image Processing 7(7) (1998)
8. Jayaraman, S., et al.: Digital Image Processing. Tata McGraw Hill Education Pvt. Ltd (2009)
9. Moertini, V.S.: Introduction to five data clustering algorithms. Integral 7(2) (2002)

Can Ear and Soft-Biometric Traits Assist in Recognition of Newborn?

Shrikant Tiwari, Aruni Singh, and Sanjay Kumar Singh

Department of Computer Engineering
Institute of Technology
Banaras Hindu University
Varanasi, India-221005
shrikant.rs.cse@itbhu.ac.in,
arunisingh@rocketmail.com,
sks.cse@itbhu.ac.in

Abstract. Missing, swapping, mixing, and illegal adoption of newborns is a global challenge and research done to solve this issue is minimal and least reported in the literature. Most of the biometric systems developed are for adults and very few of them address the issue of newborn identification.The ear of newborn is a perfect source of data for passive identification of newborn as they are the highly non cooperative users of biometrics. The four characteristics of ear biometrics: universality, uniqueness, permanence and collectability make it a very potential biometric trait for the identification of newborn. Further the use of soft-biometric data like gender, blood group, height and weight along with ear enhances the accuracy for identification of newborn. The objective of this paper is to demonstrate the concept of using ear and soft-biometrics recognition for identification of newborn. The main contribution of the research are (a) the preparation of ear and soft biometric database of newborn. (b)Fusion of ear and soft-biometrics data for identification of 210 newborn, results in an improvement of approximately 5.59% over the primary biometric system i.e. ear.

Keywords: Ear, Soft-biometric, Newborn, Recognition, Fusion.

1 Introduction

The problem of missing children is a very serious issue throughout the world and seeing the importance of this issue, May 25 is observed as National Missing Children's Day. Reliability and efficiency for newborn recognition are key to the stringent security requirements to control mixing, swapping, kidnapping and illegal adoption of newborn. The level of security is very crucial issue in maternity ward and the problem of missing and swapping of newborn is of prime concern to the persons involved and affected. There is a common perception in the society that they can do nothing to prevent this unfortunate event. In comparison to developed nations the developing countries are facing more challenges because of overcrowding and scarcity of medical facilities in the hospital. According to report every year around

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 179–192.
springerlink.com

1,00,000 *to* 5,00,000 newborns in United States are exchanged (swapped) by mistake, or one out of every eight babies born in American hospitals sent home with the wrong parents [1]. According to study, out of 34 newborns that are admitted to a neonatal intensive care unit there are 50% probabilities of incorrect newborns identification only in a single day [2, 3]. These are the number of cases that have been reported, but there may be many more cases that are undeclared or the parents and the children never come to know about this unfortunate incident.

The prime concern is that how the parents can be assured that their new born will not be mixed up in hospital. The technique of the identification procedure explained to identify newborn, hangs the peace of mind of the parents until such time as the newborn shows unmistakable evidences of its parentage.

Hospitals have devised different procedures to ensure that babies are correctly recognized and one of the popular methods is the use of ID bracelets. Soon after the birth ID bracelets are put on babies hands/legs, but this has not been able to provide enough level of security for newborn. The medical technique like Deoxyribonucleic Acid (DNA) typing and Human Leukocyte Antigen (HLA) typing are very efficient and accurate methods for verifying the identity of babies but due to the amount of time it takes to process a DNA or HLA sample and the cost associated with it, these methods for verification are not feasible for every newborn. Another method recommended by Federal Bureau of Investigation is foot and finger printing of the child and mother [4]. According to survey report 90% of the hospitals in United States perform foot printing of the babies within 2 hours of their birth and hospitals maintain newborn identification form on which footprint of the child and fingerprint of the mother are collected. The prints are generally collected using ink based methods and then printed on the identification form. Medical and computer scientist have explored the efficiency and authenticity of using footprints for newborn identification and analysis done by Shepard *et al.* using footprints of 51 newborns was examined by fingerprint experts ant they were able to identified only 10 newborn [5, 6].

Pela et. al. conducted the study on 1917 foot prints collected by trained staff of hospital in Brazil. Most of the images collected provided insufficient information for identification of newborn [7]. The American Academy of Pediatrics and others concluded that individual hospitals may continue the practice of foot printing or fingerprinting, but universal application of this practice is not recommended. After footprint, researchers explored the applicability of other biometric modalities such as fingerprint, palm print and ear for verifying the identity of newborn babies [8]. Although fingerprint and palm print recognition are well established modalities to recognize adults (over the age of 5 years), they did not achieve good results in identifying newborns. Weingaertner et al. developed a new high resolution sensor for capturing the foot and palm prints of babies [9]. Two images of 106 newborns were collected: one within 24 hours of birth and another at around 48 hours. Fingerprint experts examined the data and the identification accuracy of 67.7% and 83% were obtained using foot prints and palm prints respectively. However, multiple studies have quoted that capturing finger/palm/footprint of newborns is very challenging as it is difficult to hold their hands and legs still. Fields et al. have studied the feasibility of ear recognition on a database of 206 newborns [10]. They manually analyzed the samples and concluded that visually ears can be used to distinguish between two

children. In all the methods for identifying newborns, no research has evaluated the performance of automatic identification or verification.

Another biometric modalities that have been extensively studied for adults are face [11] and iris [12] recognition. Although iris recognition for adults yields very high accuracy [12], for newborns, it is very difficult to capture iris patterns. The work done on face recognition of newborn reports the accuracy of 86.9% on the database of 34 babies also suffers from facial expression of newborn as the face database consist of crying or sleeping face because it is very difficult to get the normal face [13]. Recently the work done by Rubisley P Lemes et al. demonstrate the use of palmprint using high resolution scanner on the database of 250 newborn has the limitation of good quality image and high cost recognition [14].

2 Why Ear and Soft Biometrics for Newborn?

Ears have gained attention in biometrics due to robustness of the ear shape [15, 16, 17, 18, 19, 20]. They have distinguishing and stable feature that changes little with age. The limitation of face biometrics compared to ear is that it does not suffer from changes in facial expression. In case of newborn ear biometric is better than other biometrics specially face because most of the time newborn are either sleeping or crying and thus it is not affected by any expression. In comparison to biometric trait like iris, retina and fingerprint ear is bigger are size and the capturing of image can be easily done at a distance.

It is our assertion that ear recognition can be a hygienic, friendly and cost effective solution for identifying newborns if the performance of automatic matching algorithms is satisfactory. In this research, we have investigated the applicability and performance of ear recognition to prevent newborn switching, illegal adoption and abduction.

Soft Biometrics characteristics like gender, blood-group, age, height, weight and head print are not unique and reliable but they provide some useful information about the individual and these are referred as soft biometric trait and these trait compliment the primary biometric trait [21,22, 23, 24]. Soft biometric traits help in filtering large databases by reducing the number of search for each query. In case of newborns we have collected gender, height, weight and blood group as soft-biometric data.

3 Database Acquisition

One of the main reasons of limited research for newborn identification is the non availability of reference database in public domain. The biggest problem in preparing the database of newborn is the consent of parents and the cooperation of medical staff to prepare the database. The active participation of parents and the medical staff provides an additional advantage for the successful preparation of the newborn database. It is really difficult to convince the parents for data acquisition as some parents were unwilling and concerned about the privacy issue. New born are highly

uncooperative users of biometrics and most of the time they are sleeping or crying. Therefore, to capture their image is really a difficult task because as soon as they are targeted for data acquisition they get disturbed and start crying. During biometric data acquisition a crucial problem faced by the biometric researchers is to decide an opportune time of the data acquisition. If a newborn is uncomfortable due to hunger or medical illness then he/she will cry and ceaselessly move his/her head, feet or whole body. Even if they are sleeping, then the task of their data acquisition becomes more challenging.

The newborns database consists of static digital images of ear (Digital camera of 10 megapixels and video camera of 14 megapixels to capture the images of ear) and soft biometrics data like gender, height, weight and blood group. The data base acquisition of newborns took one year to complete and thus it has minor variations in pose, illumination. The datasbase of newborn includes 2100 images of ear from 210 subjects with 10 images per person out of which first 4 images of each newborn is randomly selected for training/gallery database (total of 840 images) and the remaining 6 images of each newborn is selected for testing/probe database (total 1260 images). In the pre-processing step the ear part is manually cropped color images to a size of $1402\,x\,1900$ pixels as shown in Fig. 1(b). The cropped color image is converted to a gray scale image as shown in Fig. 1(c). The normalization of ear image is done in two stages. The geometric normalization scales all the images to the standard size of $160\,x\,160$. In photometric normalization different levels of masking are experimented for finding the best one to get as good accuracy as possible for the algorithm.

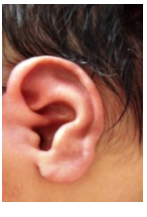


Fig. 1. (a) Before Cropping

Fig. 1. (b) After Cropping

Fig. 1. (c) Normalized Ear Image

The weight of an newborn is measured by digital weighting machine at the place where the newborn lie while providing the primary biometric. The height can be estimated from Newbornometer obtained when the newborn coming for checkup. The Fig.2 displays a mechanism for capturing of the height and weight information of an newborn. These images are captured without imposing any constraint on the newborn or their surroundings. Hence collected database is combination of pose, expression and certain illumination variations due to the newborns movement, some instances of motion blurriness also present in the newborn database. The format of soft biometric data is shown in the TABLE-I.

Table 1. Database statistics of soft biometrics

Gender Distribution	**Male**				**Female**			
	70				140			
Blood Group	A+	A-	B+	B-	AB+	AB-	O+	O-
	32	21	30	20	25	18	58	6
Height	40cm to 45 cm			46cm to 50 cm		more than 51 cm		
	50			130		30		
Weight	1500gm to 2500 gm			2501gm to 4000 gm		more than 4001gm		
	50			130		30		

Fig. 2. (a) Newbornometer

Fig. 2. (b) weighting machine

4 Covariates of Newborn Ear Database

The ear database of newborn consists of pose, illumination and occlusion covariates. Occlusion is due to some tradition that soon after birth parents put black earrings or black threads in the ear. Ear images are grouped according to variations mentioned above to solve the problem of ear recognition in newborn as shown in Fig.3 (a), 3(b) and 3(c).

Ear recognition is a long studied problem and several challenges have been identified by the researchers including pose, illumination, and occlusion. Since it is difficult to make the newborns sit still and give good ear images, they can be considered as uncooperative users of biometric recognition. They may also exhibit different poses, especially if they become uncomfortable while capturing the ear image. In some cases occlusion is also an important issue because soon after the birth some parents put black thread or ear ring due to their tradition.

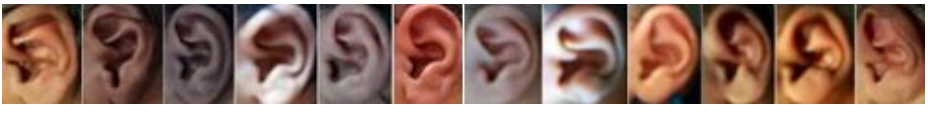


Fig. 3 (a). Illumination variation ear images of newborn from the database

Fig. 3 (b). Pose variation ear images of newborn from the database

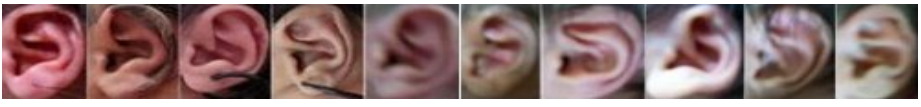


Fig. 3 (c). Occlusion variation ear images of newborn from the database

5 Frame Work for Fusion of Ear and Soft Biometric Information

In the proposed framework, the biometric recognition system is divided into two subsystems. The two subsystems are the primary biometric system which consist of ear and the secondary biometric system consisting of soft biometric traits like height, weight, gender and blood-group. Fig. 4 shows the architecture of a personal recognition system that makes use of both ear and soft biometric measurements [25].

Let $x = [\omega_1, \omega_2, \ldots, \omega_n]$ where the total no of newborns enrolled is n and x is the feature vector corresponding to the ear. The primary biometric system output is the form of $P(\omega_i|x), i = 1,2, \ldots, n,$ where $P(\omega_i|x)$ is the probability, x is the feature vector for the test user ω_i . The primary biometric system output is a matching score,

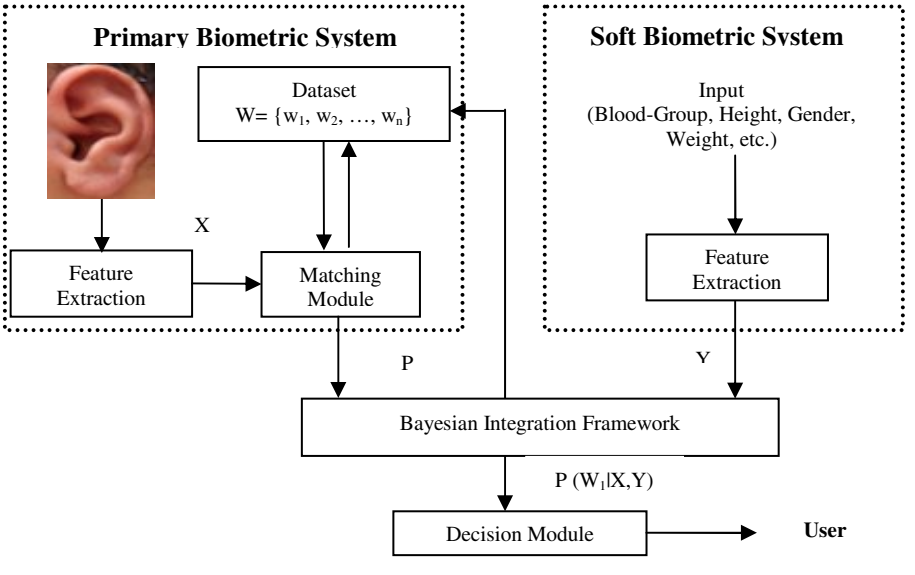


Fig. 4. Framework for fusion of primary and soft biometric information. Here x is the primary (ear) feature vector and y is the soft biometric feature vector.

which is converted into posteriori probability. For the secondary biometric system, we can consider $P(\omega_i|x)$ as the prior probability for the corresponding test user ω_i.

Let $\boldsymbol{y} = [y_1, y_2, \ldots \ldots y_k, y_{k+1}, y_{k+2}, \ldots\ldots, y_m]$ be the feature vector of soft biometric, where, y_1 through y_k denotes continuous variables and y_{k+1} through y_m are discrete variables. Finally the matching probability of the user ω_i, and the given primary biometric and soft biometrics feature vector is $\boldsymbol{x}$ and $\boldsymbol{y}$, i.e., $P(\omega_i|x, y)$ can be calculated using the Bayes' rule as[25,26,27]:

$$P(\omega_i|\boldsymbol{x}, \boldsymbol{y}) = \frac{p(y\ |\omega_i)P(\omega_i|x)}{\sum_{i=1}^{n} p(y\ |\omega_i)P(\omega_i|x)} \tag{1}$$

If the variables are independent then equation (1) can be rewritten as follows

$$P(\omega_i|\ \boldsymbol{x}, \boldsymbol{y}) = \frac{p(y_1|\ \omega_i) \ldots p(y_k|\ \omega_i)\ P(y_{k+1}|\ \omega_i) \ldots\ldots p(y_m|\ \omega_i)P(\omega_i|\ \boldsymbol{x})}{\sum_{i=1}^{n} p(y_1|\ \omega_i) \ldots p(y_k|\ \omega_i)\ P(y_{k+1}|\ \omega_i)..p(y_m|\ \omega_i)P(\omega_i|\ \boldsymbol{x})} \tag{2}$$

In equation (2), $p(y_j|\ \omega_i), j = 1,2, \ldots, k$ represents the conditional probability of the continuous variable y_j for the corresponding user ω_i. This can be evaluated from the conditional density of the variable j for the user ω_i. On the other hand, discrete probabilities $p(y_j|\ \omega_i), j = k+1, k+2, \ldots, m$ represents the probability that user ω_i is assigned to the class y_j. This is a measure of the accuracy of the classification module in assigning user ω_i to one of the distinct classes based on biometric indicator y_j.

The logarithm of $P(\omega_i|\ \boldsymbol{x}, \boldsymbol{y})$ in equation (2) can be expressed as

$$\begin{aligned} logP(\omega_i|\ \boldsymbol{x}, \boldsymbol{y}) = {} & \log p(y_1|\ \omega_i) + \cdots + \log p(y_k|\ \omega_i) \\ & + \log P(y_{k+1}|\ \omega_i) + \ldots\ldots + \log P(y_m|\ \omega_i) \\ & + \log P(\omega_i|\ \boldsymbol{x}) - \log p(\boldsymbol{y}) \end{aligned} \tag{3}$$

where $p(y) = \sum_{i=1}^{n} p(y_1|\ \omega_i) \ldots p(y_k|\ \omega_i)P(y_{k+1}|\ \omega_i) \ldots. p(y_m|\ \omega_i)P(\omega_i|\ \boldsymbol{x})$.

The resultant weight in the following discriminant function for newborn ω_i as [25]:

$$\begin{aligned} \mathrm{g_i}(\mathbf{x}, \mathbf{y}) = {} & \mathrm{a_0} \log P(\omega_i|\ \mathbf{x}) + \mathrm{a_1} \log p(y_1|\ \omega_i) \\ & + \ldots\ldots + \mathrm{a_k} \log p(y_k|\ \omega_i) + \mathrm{a_{k+1}} \log P(y_{k+1}|\ \omega_i) \\ & + \cdots\ldots + \mathrm{a_m} \log p(y_m|\ \omega_i) \end{aligned} \tag{4}$$

where $\sum_{i=0}^{m} \mathrm{a_i} = 1$ and $\mathrm{a_0} \gg \mathrm{a_i}, \mathrm{i} = 1,2, \ldots \mathrm{m}$.

Note: - For the soft biometric traits and primary biometric identifier assigned weights are the $\mathrm{a_i}$'s, $\mathrm{i} = 1,2, \ldots \mathrm{m}$ and a_0 respectively.

6 Experimental Work

In order to achieve our goal to extract features from ear we evaluate well-known, classical algorithms: PCA, KPCA, FLDA, ICA and GF.

- Principal Component Analysis (PCA) [28,29,30]

- Kernel Principal Component Analysis (KPCA) [31]
- Fisher Linear Discriminant Analysis (FLDA) [32,33,34]
- Independent Component Analysis (ICA) [35,36]
- Geometrical Feature Extraction (GF) [37, 38]

Evaluation process is performed five times for checking validation and computed rank-1 identification accuracies. The overall performance evaluations of all the five algorithms (PCA, KPCA, FLDA, ICA and GF) are computed on the newborn ear database. The results of this experiment are compiled in following TABLE-II and Fig. 5, it is observed that the identification accuracy of GF is 83.67% at Rank-I.The key analyses of the ear recognition are explained below

- The difficulty of ear feature extraction lies in the changes among the same ear caused by head rotation and lighting variation because most of the time newborn are sleeping or crying. The geometry feature extraction depends heavily on the quality of the image preprocessing.
- Due to different lighting conditions the curve segments extraction and the structural extraction will be different even for the same newly born child, which makes the methods unreliable. The rotation discrimination is even more challenging because the angle between the ear and the head is not the same among different babies.
- TABLE II shows that among the appearance based algorithms, FLDA provides the best accuracy of 80.57% at the Rank-1Level. The performance of appearance based PCA, KPCA, FLDA and ICA algorithm increase with increasing the levels of Gaussian pyramid i.e. decreasing the resolution of the image.
- For Geometrical Feature Extraction (GF) method works on the concept of finding out the points on the contour and distance between them, so the result is approximated in our algorithm by allowing an error of 2% and accuracy is 83.67%.
- Through experiment, we found that recognition performance of appearance methods (such as PCA, KPCA, FLDA and ICA) will increase dramatically when the input image contains much less background information around the ear.

Table 2. Identification accuracy of the newborn database

Procedure	Identification Accuracy (Rank-1)
PCA	78.56
KPCA	80.03
ICA	71.75
FLDA	80.57
GF	83.67

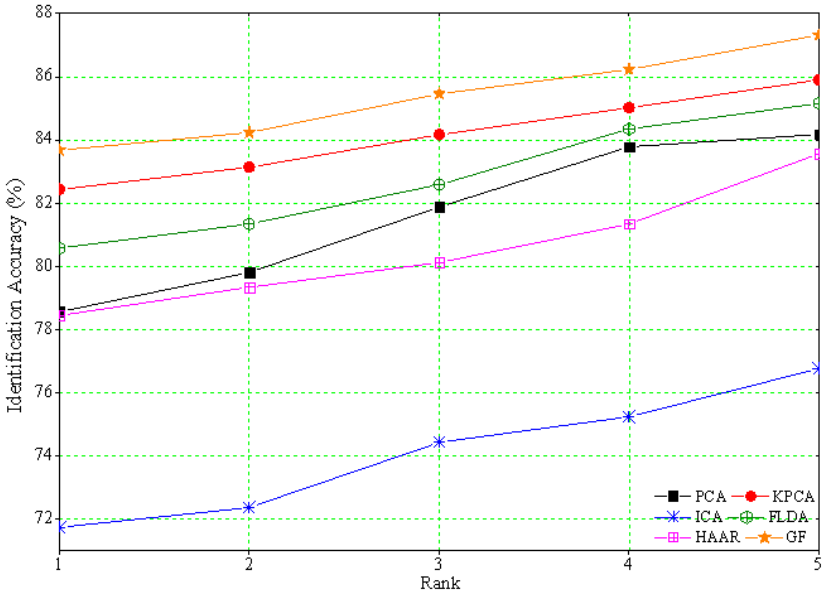


Fig. 5. CMC for ear recognition algorithm

7 Performance Gain by Fusion of Ear and Soft-Biometric

In our experiments we have selected soft biometric traits such as gender, blood group, weight, and height information of the user in addition to the ear biometric identifiers.

Let $P(\omega_i|s)$ be the posterior probability (Ear) that the user is newborn $\omega_i$given the primary biometric score 's' of the test user. Let $y_i = (G_i, B_i, W_i, H_i)$ is the soft biometric feature vector corresponding to the identity claimed by the user ω_i , where G_i, B_i, W_i and H_i are the true values of gender, blood group, weight, and height of ω_i. Let $y^* = (G^*; B^*; W^*; H^*)$ is the soft biometric feature vector of the observed test user, where G^* is the observed gender, B^* is the observed blood group, W^* is the observed weight, and H^* is the observed height. Finally the score after considering the observed soft biometric characteristics is computed as

$$g_i(\mathbf{s}, y^*) = a_0 \log P(genuine|s) + a_1 \log p(H^*|H_i) + a_2 \log P(W^*|W_i) + a_3 \log P(G^*|G_i) + a_4 \log P(B^*|B_i)$$

where $a_3 = 0$,. if G^* ="reject", and $a_4 = 0$ if B^* ="reject".

Fig.5 shows the Cumulative Match Characteristic (CMC) of the ear biometric system operating in the identification mode, and the improvement in performance achieved after the utilization of soft biometric information. The weights assigned to the ear (primary) and soft biometric traits were selected experimentally such that the performance gain is maximized. However, no formal procedure was used and an exhaustive search of all possible sets of weights was not attempted. The use of blood-group, height, weight and gender information along with the ear leads to an improvement of 5.59% in the rank one performance as shown in Fig. 6(a), 6(b), 6(c)

and Fig. 6(d) respectively. From 6(b), 6(c) and Fig. 6(d), we can observe that the blood-group information of the newborn is more discriminative than gender and leads to a 1.49% improvement in the rank one performance. The combined use of all the four soft biometric traits results in an improvement of approximately 5.59% over the primary biometric system as shown in Fig. 6(e).

Table 3. Identification accuracy of the newborn database

Procedure	**E**	**E + G**	**E + H**	**E +W**	**E + B**	**E +G+H+W+B**
Identification Accuracy (Rank-1)	83.67	85.12	86.46	86.16	85.16	89.26

where E=Ear, G=Gender, H=Height, W=Weight, B=Blood-group

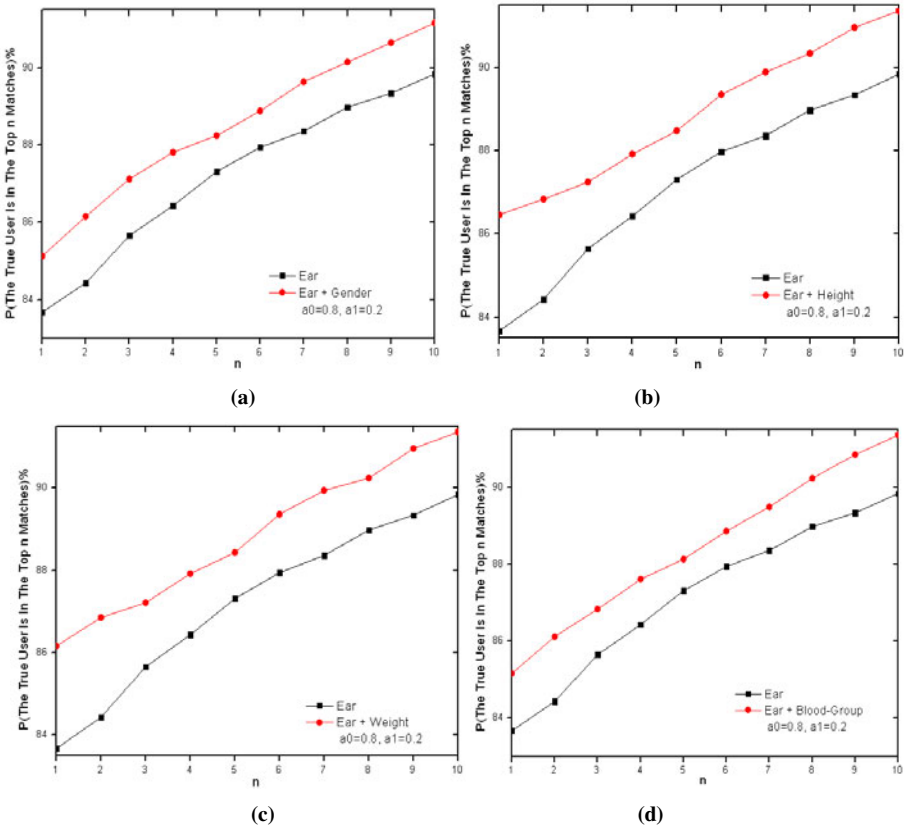
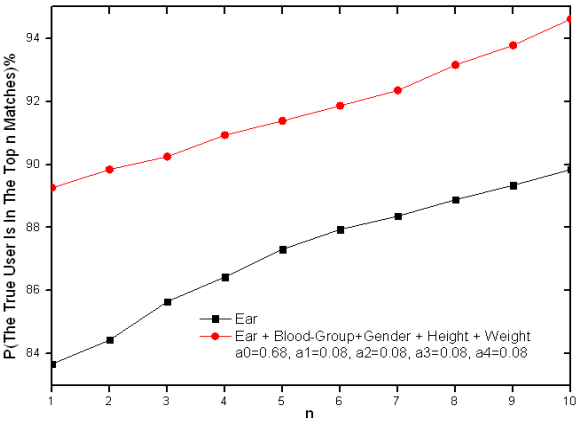


Fig. 6. Improvement in the performance of a unimodal (Ear) system after addition of soft biometric traits, (a) Ear with Gender, b) Ear with Height, c) Ear with Weight, d) Ear with Blood-Group and e) Ear with Blood-Group ,Gender, Height and Weight



(e)

Fig. 6. (*Continued*)

8 Conclusion and Future Direction

Mixing and kidnapping of newborn is a strong negative response, many parents fear that there is nothing they can do to prevent this tragedy. There is a lot of justification for recognition of newborns using biometrics to mitigate the problem of mixing, switching, abduction and some of the biometric traits collected in the prepared database are justified for only some limited time duration.

The objective of this paper is to demonstrate that ear and soft biometric identifiers such as height, weight, gender, and blood-group can be very useful in newborn recognition. It is our assertion that ear and soft biometric data can be a very promising tool for identification of newborn. Although the soft biometric characteristics are not as permanent and reliable as the traditional biometric identifiers like ear, they provide some information about the identity of the newborn that leads to higher accuracy in establishing the user identity.

Our proposed model demonstrated that the utilization of ancillary user information like gender, height, weight and blood-group can improve the performance of the traditional biometric system. Although the soft biometric characteristic are not as permanent and reliable as the traditional biometric identifiers like ear, they provide some information about the identity of the user that leads to higher accuracy in establishing the user identity.

The approach described in this paper is relatively successful and promising in ear recognition of newborn with fusion of soft biometric data, but more research is to be done by the scientist and engineers in the following domain.

- Size of database is to be increased and following conditions may be considered while capturing images of each subject: illumination, variation, pose, variation, distance variation, date-variation and occlusion variation.
- Collections of ear image of newborn after certain interval of time and then analyze the efficacy of ear recognition in newborn.

- Design and development of pose invariant algorithms as pose is an important covariate in newborn because newborn are highly uncooperative user of biometrics.
- Illumination is also a big challenge because of changing weather condition and the location (indoor or outdoor). So an illumination invariant technique is to be developed.
- In case of newborn occlusion is also a problem as in some case it is found that soon after the birth the ear is pierced and black thread is inserted and some ears are found to be infected by some disease.
- Ear and Face multimodal biometrics can be used to enhance the identification accuracy and security level.
- To make the enrollment process automatic there is a need to construct a model of variation.

Acknowledgement. Authors would like to thank Prof. B. M. Singh and Dr. Niraj Srivastava (Department of Kaumar Bharitya, Faculty of Aurveda, Institute of Medical Science, BHU, Varanasi-India) for their help and cooperation in preparing the database.

References

1. http://www.amfor.net/stolenbabies.html (last accessed on May 25, 2011)
2. http://www.missingkids.com/enus/documents/infantabductionstats.pdf (last accessed on June 4, 2011)
3. Gray, J.E., Suresh, G., Ursprung, R., Edwards, W.H., Nickerson, J., Shinno, P.H.: Patient Misidentification in the neonatal intensive care unit: Quantification of ris. Paediatrics 117, e46–e47 (2006)
4. Stapleton, M.E.: Best foot forward: Infant footprints for personal identification. Law Enforcement Bulletin 63, FBI (1999)
5. Shepard, K.S., Erickson, T., Fromm, H.: Limitations of footprinting as a means of infant identification. Pediatrics 37(1) (1996)
6. Thompson, J.E., Clark, D.A., Salisbury, B., Cahill, J.: Footprinting the infant: not cost-effective. Journal of Pediatrics, 797–798 (1981)
7. Pela, N.T.R., Mamede, M.V., Tavares, M.S.G.: Analise crıtica de impressoes plantares de recem-nascidos. Revista Brasileira de Enfermagem, 100–105 (1975)
8. Galton, F.: Finger prints of young children. British Association for the Advancement of Science (1989)
9. Weingaertner, D., Bellon, O.R.P., Cat, M.N.L., Silva, L.: Infant's biometric identification: Can it be done? In: International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (2008)
10. Fields, C., Hugh, C.F., Warren, C.P., Zimberoff, M.: The ear of the infant as an identification constant. Journal of Obstetrics and Gynecology 16, 98–101 (1960)
11. Li, S.Z., Jain, A.K.: Handbook of Face Recognition. Springer, New York (2004)
12. Daugman, J.: New methods in iris recognition. IEEE Transactions on Systems, Man and Cybernetics B 37(5), 1167–1175 (2007)

13. Bharadwaj, S., Bhatt, H.S., Singh, R., Vatsa, M., Singh, S.K.: Face Recognition for Infants: A Preliminary Study. In: Fourth IEEE International Conference on Biometrics, Theory Applications and Systems (BTAS), pp. 1–6, 27–29 (2010)
14. Lemes, R.P., Bellon, O.R.P., Silva, L., Jain, A.K.: Biometric Recognition of Newborns: Identification using Palmprints. In: International Joint Conference on Biometrics, Washington DC, USA, October 11-13 (2011)
15. Kuefner, D., Cassia, V.M., Picozzi, M., Bricolo, E.: Do all kids look alike? evidence for another-age effect in adults. Journal of Experimental Psychology: Human Perception and Performance 34(4), 811–817 (2008)
16. Lomuto, C., Duverges, C.: Identificacion delrecien nacidoy medidas de prevencion para evitar surobo delas maternidades. Revista del Hospital Materno Infantil Ramon Sarda 14(3), 115–124 (1995)
17. Iannarelli, A.: Ear Identification. Paramont Publishing Company (1989)
18. Pun, K.H., Moon, Y.S.: Recent advances in ear biometrics. In: Proceedings of the Sixth International Conference on Automatic Face and Gesture Recognition, pp. 164–169 (2004)
19. Yuizono, T., Wang, Y., Satoh, K., Nakayama, S.: Study on individual recognition for ear images by using genetic local search. In: Proceedings of the 2002 Congress on Evolutionary Computation, pp. 237–242 (2002)
20. Burge, M., Burger, W.: Ear biometrics in Computer Vision. In: International Conference of Pattern Recognition, pp. 822–826 (2000)
21. Ross, A., Nandakumar, K., Jain, A.K.: Handbook of Multibiometrics. Springer, New York (2006)
22. Jain, A.K., Dass, S.C., Nandakumar, K.: Can soft biometric traits assist user recognition? In: Proceedings of SPIE International Symposium on Defense and Security: Biometric Technology for Human Identification (2004)
23. Gutta, S., Huang, J.R.J., Jonathon, P., Wechsler, H.: Mixture of Experts for Classification of Gender, Ethnic Origin, and Pose of Human Faces. IEEE Transactions on Neural Networks 11, 948–960 (2000)
24. Jain, A.K., Dass, S.C., Nandakumar, K.: Soft Biometric Traits for Personal Recognition Systems. In: Zhang, D., Jain, A.K. (eds.) ICBA 2004. LNCS, vol. 3072, pp. 731–738. Springer, Heidelberg (2004)
25. Jain, A.K., Nandakumar, K., Lu, X., Park, U.: Integrating Faces, Fingerprints, and Soft Biometric Traits for User Recognition. In: Maltoni, D., Jain, A.K. (eds.) BioAW 2004. LNCS, vol. 3087, pp. 259–269. Springer, Heidelberg (2004)
26. Jain, A.K., Lu, X.: Ethnicity Identification from Face Image. In: Proceedings of SPIE International Symposium on Defense and Security: Biometric Technology for Human Identification (2004)
27. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley & Sons (2001)
28. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 711–720 (1997)
29. Song, Y.-J., Kim, Y.-G., Kim, N., Ahn, J.-H.: Face Recognition using both Geometric Features and PCA/LDA. In: Sixth International Conference on Advanced Language Processing and Web Information Technology
30. Ping, Y., Bowyer, K.W.: Empirical Evaluation of Advanced Ear Biometrics. In: Proc. Empirical Evaluation Methods in Computer Vision, San Diego, pp. 56–59 (2005)
31. Li, Y.: Study on Some Key Issues in Ear Recognition. PhD thesis, University of Science and Technology Beijing, Beijing (2006)

32. Kurita, T., Taguchi, T.: A Modification of Kernel-based Fisher Discriminant Analysis for Face Detection
33. Liu, W., Wang, Y., Li, S.Z., Tan, T.: Null Space Approach of Fisher Discriminant Analysis for Face Recognition. In: Proceeding of ECCV Workshop on Biometric Authentication, pp. 32–44 (2004)
34. Chen, L.F., Mark Liao, H.Y., Ko, M.T., Yu, G.J.: A New LDA-based Face Recognition System Which Can Solve the Small Size Problem. Pattern Recognition 33(10), 1713–1726 (2000)
35. Bartlett, M.S., Movellan, J.R., Sejnowski, T.J.: Face recognition by independent component analysis. IEEE Transactions on Neural Networks 13(6), 1450–1464 (2002)
36. Nanni, L., Lumini, A.: A multi-matcher for ear authentication. Pattern Recognit. Lett. 28(16), 2219–2226 (2007)
37. Choras, M.: Ear biometrics based on geometrical features extraction. Electron. Lett. Comput. Vis. Image Anal. 5(3), 84–95 (2005)
38. Choraś, M., Choraś, R.S.: Geometrical Algorithms of Ear Contour Shape Representation and Feature Extraction. In: Proc. of Intelligent Systems Design and Applications (ISDA), Jinan, China, vol. II, pp. 451–456. IEEE CS Press (2006)

Multi Segment Histogram Equalization for Brightness Preserving Contrast Enhancement

Mohd. Farhan Khan, Ekram Khan, and Z.A. Abbasi

Department of Electronics Engineering, Aligarh Muslim University, Aligarh
{farhan7787,ekhan67}@gmail.com, ziaabbasi@rediffmail.com

Abstract. Histogram equalization (HE) method proved to be a simple and most effective technique for contrast enhancement of digital images, but it does not preserve the brightness and natural look of images. To overcome this problem, several Bi- and Multi-histogram equalization methods have been proposed. Among them, the Bi-HE methods significantly enhance the contrast and may preserve the brightness, but they destroy the natural look of the image. On the other hand, Multi-HE methods are proposed to maintain the natural look of image at the cost of either the brightness or its contrast. In this paper, we propose a Multi-HE method for contrast enhancement of natural images while preserving its brightness and natural look. The proposed method decomposes the histogram of an input image into multiple segments, and then HE is applied to each segment independently. Simulation results for several test images show that the proposed method enhances the contrast while preserving brightness and natural look of the images.

Keywords: Histogram equalization, histogram segmentation, contrast enhancement, brightness preserving.

1 Introduction

Image enhancement process involves mapping the pixel's intensity of the input image, so that the processed image should subjectively looks better [3]. Many image enhancement methods have been developed. A very popular and most effective technique for image enhancement is histogram equalization (HE). HE becomes popular for contrast enhancement because of its simplicity and effectiveness. Its basic idea lies on mapping the gray levels based on the probability distribution of the input gray levels. It flattens and stretches the dynamics range of the image's histogram and resulting in overall contrast enhancement. HE has been successfully applied in various fields such as medical and satellite image processing [5]. Despite of its popularity, HE is not very suitable to be implemented in consumer electronics, such as video surveillance, due to fact that HE normally shifts the brightness of the input image significantly. Thus, for contrast enhancement in the consumer electronic products, it is advisable that the processed image retain the brightness and natural look of the input image [11].

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 193–202.
springerlink.com

Many methods with multiple histogram equalization have been proposed to achieve this objective [8]. The earliest work in Bi-HE area has been reported by Kim in [5], with a technique known as brightness preserving bi-histogram equalization (BBHE). It is mathematically shown that the mean brightness of processed image by BBHE method, locates in the middle of the input mean and the middle gray level (i.e., L/2) [5]. Then, dualistic sub-image histogram equalization (DSIHE) [3] has been proposed by Wan et al. The mean brightness of the processed image by DSIHE method is at the average of the segmented gray level and the middle gray level [11]. Minimum mean brightness error bi-histogram equalization (MMBEBHE) is then proposed by Chen and Ramli to "optimally" maintain the mean brightness. MMBEBHE method first test's all the possible values of the separating intensity, from 0 to L-1, and then the differences between the mean brightness of the input image and the mean brightness of the processed images are calculated. Then, the value of threshold is chosen, by enumeration, as the value that can produce the minimum different between input and output means [2]. Bi-HE methods usually destroy natural appearance of the image [6].

To preserve the natural look of the image, Multi-HE method was introduced, which prevents the shift of mean brightness in processed images as compared to input images but may not significantly enhance the contrast. Chen and Ramli proposed recursive mean-separate histogram equalization (RMSHE) method. It is claimed that RMSHE is good brightness preservation technique when the value of r is large [1]. However, it can be observed that when r is too large, the output histogram will become same as input histogram. Hence, the processed image is exactly the copy of the input image with no contrast enhancement [10]. Sim et al [9] propose a similar method to RMSHE known as recursive sub-image histogram equalization (RSIHE). Minimum Within-Class Variance Multi Histogram Equalization (MWCVMHE) and Minimum Middle Level Squared Error Multi Histogram Equalization (MMLSEMHE) are proposed by D. Menotti et al. in 2007. MWCVMHE and MMLSEMHE method preserves the brightness to much extent but the contrast enhancement is less intensive [6].

This paper proposes a method, known as Hybrid Multi Segment Histogram Equalization (HMSHE). It is actually a hybrid form of RMSHE and MWCVMHE. The method partitions the histogram into segments based on the mean [5, 1] and, minimum within class variance [6]. Depending on input image, the selected threshold may change the mean brightness. The final step involves the normalization of the output mean. With this criterion, HMSHE is expected to produce better contrast enhancement, and better in preserving the mean brightness.

This paper is organized as follows. Section 2 describes proposed method in detail. Simulation results and comparisons with other contemporary techniques are presented in Section 3. Section 4 presents the conclusion of this work.

2 Hybrid Multi Segment Histogram Equalization (HMSHE)

Hybrid Multi Segment Histogram Equalization (HMSHE) which is proposed in this paper consists of three steps: Detection of optimal thresholds, equalizing each segment independently, and normalization of image brightness. The details of each step are described in the following subsections.

2.1 Detection of Optimal Thresholds

In this section, the clustering of the image histogram into multiple segments is done, where each segment represents sub-histogram. The selection of optimal thresholds for segmentation depends on the mean of global histogram [5], and the minimum within sub-histogram class variance [6], where the within-class variance is the total squared error of each histogram class with respect to its mean brightness. The objective of histogram segmentation is to minimize the decomposition error of the input image histogram.

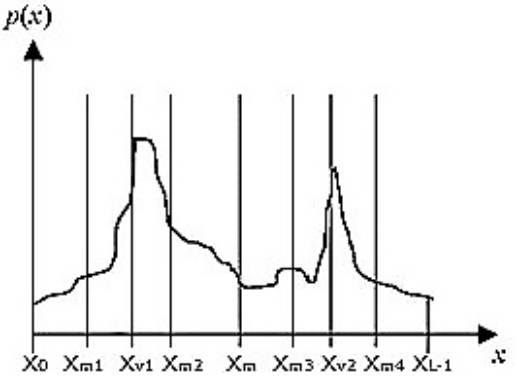


Fig. 1. Selection of optimal thresholds

Fig. 1 shows the selection of optimal thresholds before equalizing the segments of the histogram independently. The global histogram H[h_l , h_u], where l is the lowest gray level and u is upper-most gray level (for gray scale image, $l = 0$ & $u = L - 1 = 255$) is divided into n-parts (here n=8), firstly H[h_l , h_u] is segmented into H[h_l , h_m] & H[h_{m+1} , h_u] via threshold X_m, where, X_m represents the mean brightness of an image. Then again segmenting the sub-histograms by considering the global minima of within sub-histogram class variance i.e., H[h_l , h_m] is segmented in H[h_l , h_{v1}] & H[h_{v1+1} , h_m] where v_1 is the global minima of within sub-histogram class variance of H[h_l , h_m], & similarly H[h_{m+1} , h_u] is segmented in H[h_{m+1} , h_{v2}] & H[h_{v2+1} , h_u] where v_2 is the global minima of within sub-histogram class variance of H[h_{m+1} , u]. The output mean brightness [E(Y)] can be obtained as:

$$\begin{aligned}[\mathrm{E(Y)}] = {} & \mathrm{E(Y|X} \le X_{m1})P_r(\mathrm{X} \le X_{m1}) \\ & + \mathrm{E(Y|}X_{m1} < X \le X_{v1})P_r(X_{m1} < X \le X_{v1}) \\ & + \mathrm{E(Y|}X_{v1} < X \le X_{m2})P_r(X_{v1} < X \le X_{m2}) \\ & + \mathrm{E(Y|}X_{m2} < X \le X_m)P_r(X_{m2} < X \le X_m) \\ & + \mathrm{E(Y|}X_m < X \le X_{m3})P_r(X_m < X \le X_{m3}) \\ & + \mathrm{E(Y|}X_{m3} < X \le X_{v2})P_r(X_{m3} < X \le X_{v2}) \\ & + \mathrm{E(Y|}X_{v2} < X \le X_{m4})P_r(X_{v2} < X \le X_{m4}) \\ & + \mathrm{E(Y|X} > X_{m4})P_r(\mathrm{X} > X_{m4}).\end{aligned} \tag{1}$$

where, P_r is the distribution probability.

So the mean brightness of the output image of proposed method can be obtained as:

$$E(Y) = X_0 \frac{p_1}{2} + X_{m1}\left(\frac{p_1 + p_2}{2}\right) + X_{v1}\left(\frac{p_2 + p_3}{2}\right) + X_{m2}\left(\frac{p_3 + p_4}{2}\right) + X_m\left(\frac{p_4 + p_5}{2}\right) + X_{m3}\left(\frac{p_5 + p_6}{2}\right) + X_{v2}\left(\frac{p_6 + p_7}{2}\right) + X_{m4}\left(\frac{p_7 + p_8}{2}\right) + X_{L-1}\frac{p_8}{2} \quad (2)$$

Special case 1: If $p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = p_8$:
It represents RMSHE (r=3).
For above p_n =1/8, n=1, 2, .. , 8

$$E(Y) = X_m + \left[\frac{Y - X_m}{8}\right] \quad (3)$$

Special case 2: If "$X_{v1} = X_0$ or $X_{v1} = X_m$" & "$X_{v2} = X_m$ or $X_{v2} = X_{L-1}$"
It represents RMSHE (r=2).

$$E(Y) = X_m + \left[\frac{Y - X_m}{4}\right] \quad (4)$$

where, $Y = \frac{X_0 + X_{L-1}}{2}$

2.2 Equalizing Each Segment Independently

In this step, after the selection of thresholds, each sub-histogram is equalized individually by HE method. Here the output level *k*' for the input level *k* can be computed as the transformation function $T^{H[h_{lower},\ h_{upper}]}$ i.e.

$$l' = T^{H[h_{lower},\ h_{upper}]}(k) = h_{lower} + \left\lfloor (h_{upper} - h_{lower}) \times C_k^{H[h_{lower},\ h_{upper}]} \right\rfloor \quad (5)$$

where, $\lfloor f(h) \rceil$ is nearest integer function; h_{lower} and h_{upper} are lower and upper limit of sub-histograms; $C_l^{H[h_{lower},\ h_{upper}]}$ is the cumulative probability density of the level *k* in the histogram $H\left[h_{lower},\ h_{upper}\right]$.

2.3 Normalization of Image Brightness

In this section, the mean brightness of the input image, m_i, and the mean brightness of the output image obtained after the sub-histogram equalization process, m_o, is calculated. In order to normalize the shifted mean brightness of final output image $O(u, v)$

closer to the mean brightness of the input image $I(u,v)$, the brightness normalization is applied which is define as:

$$O(u,v) = \frac{m_i}{m_o} I(u,v) \tag{6}$$

3 Results and Discussion

The input images used in this paper were previously used in [1,6]. Images were extracted from the CVG-UGR-database [12]. The input image description considered for MATLAB simulation are shown in Table 1, in terms of brightness (mean), contrast (standard deviation) and dimension. In order to investigate whether the proposed method successfully maintain the input mean brightness, the results are shown in the form of AMBE & difference of output-input standard deviation. AMBE is calculated as follows:

$$\text{AMBE} = | \text{E[Y]} - \text{E[X]} | \tag{7}$$

where, E[Y] and E[X] are mean brightness of new and original gray level image, respectively [7]. So, Average Absolute Mean Brightness Error (AAMBE) is defined as follow:

$$\text{AAMBE} = \frac{1}{N}\sum_{n=1}^{N} | E_n[\text{Y}] - E_n[\text{X}] | \tag{8}$$

where, N is the total number of test images, En(X) is the average intensity of input image n, while En(Y) is the average intensity of the corresponding output image [4].

Table 1. Input image description in terms of brightness, contrast and dimension

Images	Brightness	Contrast	Dimension
copter	191.52	40.72	254x199
girl	139.25	29.64	254x254
Einstein	112.81	31.08	256x256
jet	201.17	52.02	384x256
woman	113.18	49.24	254x254
F16	179.19	45.11	512x512
hare	228.75	40.82	593x400

Table 2 defines the AMBE and Table 3 defines SD difference among various HE methods. The table data is divided into three parts: 1) The data values obtained by Uni-HE i.e. Global HE; 2) The data values obtained by Bi-HE methods i.e. BBHE and DSIHE; and 3) The data values obtained by Multi-HE methods i.e. RMSHE(r=2), MWCVMHE, MMLSEMHE and our proposed method HMSHE.

Table 2. Comparison of various methods in terms of AMBE between input and processed images

Image	Uni-HE	Bi-HE		Multi-HE			
	Global HE	BBHE [2]	DSIHE [3]	RMSHE (r=2) [6]	MWCVMHE [8]	MMLSEMHE [8]	HMSHE
copter	62.75	17.21	26.97	3.01	1.24	0.91	1.44
girl	5.26	23.58	7.54	0.57	0.26	0.85	0.21
Einstein	21.08	19.24	12.03	10.1	2.75	0.89	0.05
jet	71.78	4.96	26.86	0.69	0.36	0.59	0.07
woman	15.39	15.94	11.3	0.13	0.71	0.11	0.2
F16	49.37	1.02	15.96	1.2	5.79	0.27	0.1
hare	81.44	21.71	36.13	2.84	2.93	0.76	3.57

Let us first analyze the results in Table 2 by observing the AMBE between the original and the processed images, the images processed by Multi-HE methods preserves the mean brightness to much extent as compared to Bi-HE methods. HMSHE method is not always best among all the methods for preserving the brightness and but its resulting AMBE is closer to original images. MMLSEMHE also preserves the mean brightness of the image as compared to other Multi-HE methods. The AAMBE for HMSHE method comes out to be approx. 0.8 for seven input images considered for simulation.

Difference of output-input standard deviation (SD) defines the contrast enhancement of an image w.r.t. its corresponding input image. Higher the difference, higher will be the contrast enhancement, it is defined as:

$$\text{SD difference} = \text{SD}[\text{Y}] - \text{SD}[\text{X}] \tag{9}$$

where, SD[Y] and SD[X] are standard deviation of new and original gray level image, respectively.

Table 3. Comparison of various methods in terms of SD difference between input and processed images

Image	Uni-HE	Bi-HE		Multi-HE			
	Global HE	BBHE [2]	DSIHE [3]	RMSHE (r=2) [6]	MWCVMHE [8]	MMLSEMHE [8]	HMSHE
copter	33.24	32.04	36.1	11.39	4.19	3.82	5.43
girl	45.77	40.39	45.72	8.11	5.67	1.77	5.86
Einstein	36.45	36.76	36.81	20.81	2.94	0.48	27.15
jet	22.31	12.71	28.33	4.79	5.19	3.92	5.93
woman	24.32	24.43	24.5	13.38	3.07	1.5	11.43
F16	29.45	22.56	32.28	15.96	10.8	1.62	12.26
hare	37.66	22.84	32.22	3.69	7.62	0.04	11.45

From Table 3 it can be observed that: 1) Uni-HE and Bi-HE methods enhances the contrast to much extent as compared to Multi-HE methods; 2) The DSIHE method produces the best image contrast enhancement between Bi-HE methods; 3) The HMSHE method presents the better image contrast enhancement along with RMSHE(r=2) among the Multi-HE methods; 4) MMLSEMHE hardly enhance the contrast of the images as compared to other Multi-HE methods.

Analyzing the data of Table 2 and 3 together, it can be noticed that, Uni-HE and Bi-HE methods enhances the contrast but fails to preserve the brightness of the images. On the other hand, Multi-HE methods preserve the brightness of the image as compared to Uni- and Bi-HE methods but at the cost of its contrast. HMSHE is better than other HE methods when brightness preservation along with contrast enhancement is considered together.

Fig. 2 and 3 compares the enhancement for the copter and girl images based on Uni-HE, Bi-HE and Multi-HE methods. From Fig.2, it can be noticed that image processed by Uni-HE (i.e. Global HE), Bi-HE (i.e. BBHE, DSIHE, MMBEBHE) and

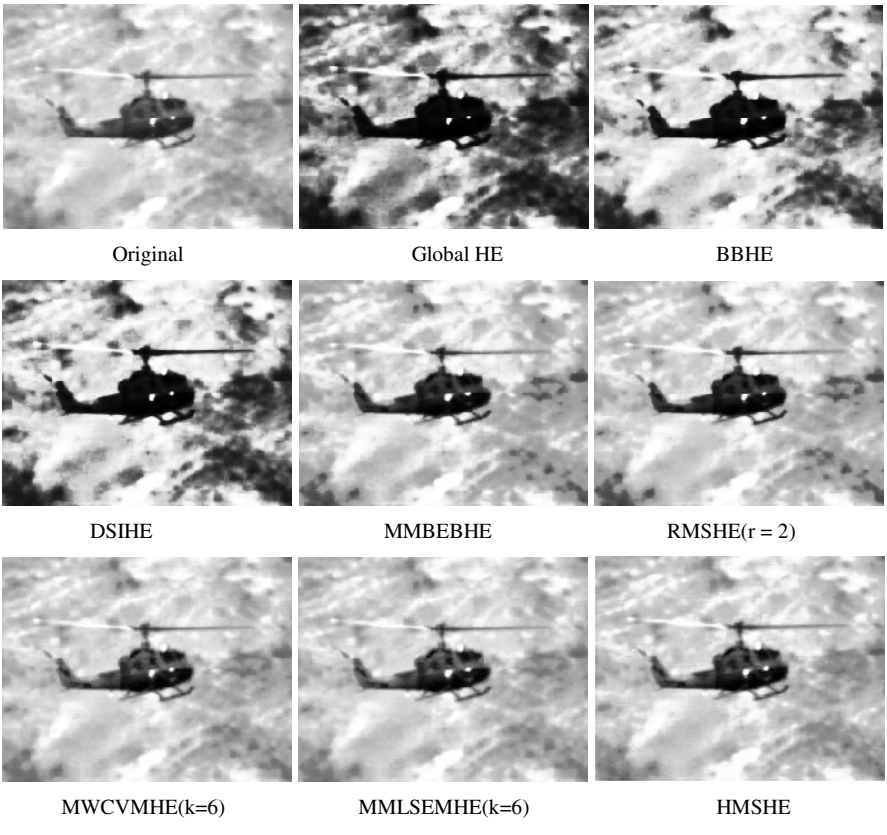


Fig. 2. Enhancement of the copter image based on Global HE, BBHE, DSIHE, MMBEBHE, RMSHE(r=2), MWCVMHE(k=6), MMLSEMHE(k=6), and HMSHE(proposed) methods.

one of the Multi-HE (i.e. RMSHE(r=2)) destroyed the natural look of the original image which can be noticed from distorted background and less perceptive details of the copter, whereas other Multi-HE methods i.e. MWCVMHE(k=6), MMLSEMHE(k=6), and HMSHE (proposed) methods preserved the natural look of the original image.

Similarly from Fig.3, it can be noticed that image processed by Uni-HE (i.e. Global HE) and Bi-HE (i.e. BBHE, DSIHE, MMBEBHE) destroyed the natural look of the original image which can be noticed from distorted background and foreground details of the image, whereas Multi-HE methods i.e. RMSHE(r=2), MWCVMHE(k=5), MMLSEMHE(k=6), and HMSHE (proposed) methods preserved the natural look of the original image.

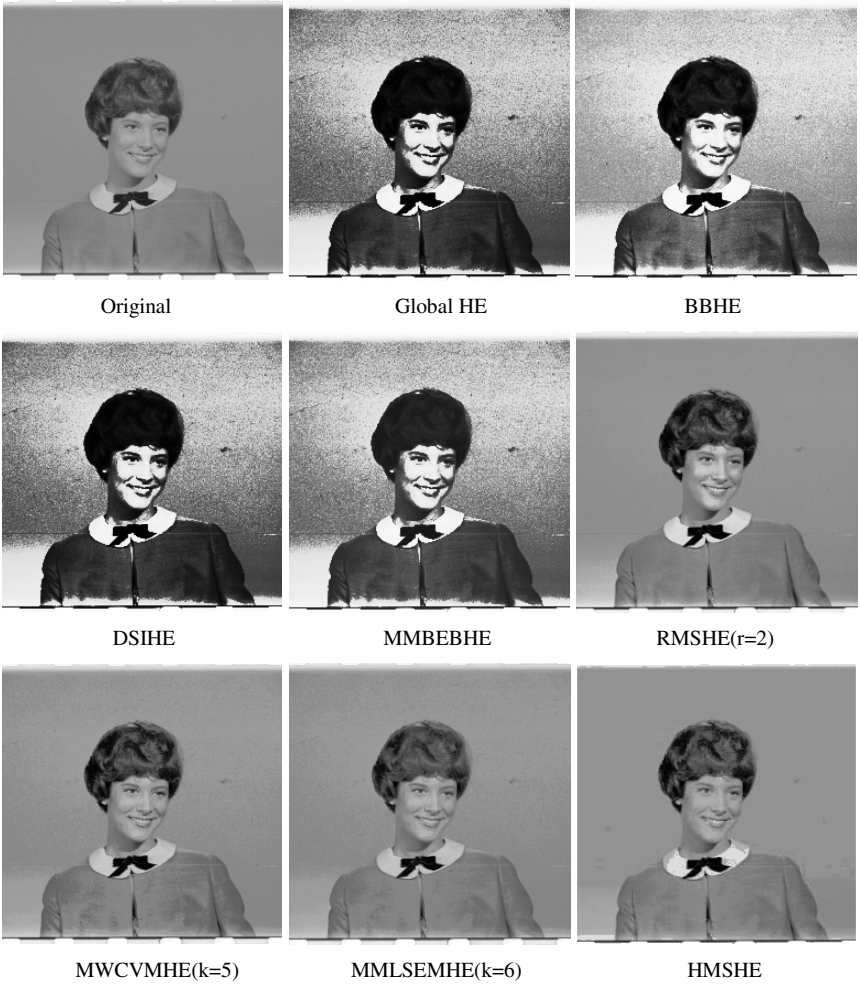


Fig. 3. Enhancement of the girl image based on Global HE, BBHE, DSIHE, MMBEBHE, RMSHE(r=2), MWCVMHE(k=5), MMLSEMHE(k=6), and HMSHE(proposed) methods

After visually observing fig. 2 and 3 and analyzing the data presented in table 2 and 3, it can be concluded that: 1) Image processed by Multi-HE methods preserves the natural look of the original image whereas, Bi-HE methods fail to preserve the natural look of the original image; 2) HMSHE method gives better results compared to other Multi-HE methods, when contrast enhancement along with preserving natural look and brightness is desired. 3) HMSHE and RMSHE(r=2) should be employed when higher contrast enhancement is desired. 4) MMLSEMHE and HMSHE should be employed when preservation of brightness along with natural look, is desired.

4 Conclusion

In this paper, HMSHE method has been proposed and tested, as a hybrid form of RMSHE and MWCVMHE. The simulation results showed that Bi-HE methods significantly enhance the contrast and may preserve the brightness, but they destroy the natural look of the image, which is undesirable in consumer electronics; while Multi-HE method may maintains the natural look of image at the cost of either the brightness or its contrast. HMSHE method is better among Multi-HE methods when contrast enhancement along with preserving the brightness and natural look of an image is desired. HMSHE is easy to implement and can be used in real time system because of its simplicity. So the advantage of proposed method is three folds.

References

1. Chen, S.D., Ramli, A.R.: Contrast enhancement using recursive mean-separate histogram equalization for scalable brightness preservation. IEEE Transactions on Consumer Electronics 49(4), 1301–1309 (2003)
2. Chen, S.D., Ramli, A.R.: Minimum mean brightness error bi-histogram equalization in contrast enhancement. IEEE Transactions on Consumer Electronics 49(4), 1310–1319 (2003)
3. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 3rd edn. Prentice Hall, India (2009)
4. Ibrahim, H., Kong, N.S.P.: Brightness Preserving Dynamic Histogram Equalization for Image Contrast Enhancement. IEEE Transactions on Consumer Electronics 53(4), 1752–1758 (2007)
5. Kim, Y.T.: Contrast Enhancement Using Brightness Preserving Bi-Histogram Equalization. IEEE Transactions on Consumer Electronics 43(1), 1–8 (1997)
6. Menotti, D., Najman, L., Facon, J., Araújo, A.A.: Multi-Histogram Equalization Methods for Contrast Enhancement and Brightness Preserving. IEEE Transactions on Consumer Electronics 53(3), 1186–1194 (2007)
7. Phanthuna, N., Cheevasuvit, F., Chitwong, S.: Contrast enhancement for minimum mean brightness error from histogram partitioning. In: Annual Conference on American Society for Photogrammetry and Remote Sensing (March 2009)
8. Rajavel, P.: Image Dependent Brightness Preserving Histogram Equalization. IEEE Transactions on Consumer Electronics 56(2), 756–763 (2010)

9. Sim, K.S., Tso, C.P., Tan, Y.Y.: Recursive sub-image histogram equalization applied to gray scale images. Pattern Recognition Letters 28(10), 1209–1221 (2007)
10. Wang, C., Ye, Z.: Brightness preserving histogram equalization with maximum entropy: a variational perspective. IEEE Transactions on Consumer Electronics 51(4), 1324–1326 (2005)
11. Wan, Y., Chen, Q., Zhang, B.M.: Image enhancement based on equal area dualistic sub-image histogram equalization method. IEEE Transactions on Consumer Electronics 45(1), 68–75 (1999)
12. CVG-URG database (2007), `http://decsai.ugr.es/cvg/dbimagenes/`

Various Implementations of Advanced Dynamic Signature Verification System

Jin Whan Kim

Dept. of Computer Engineering, Youngsan University, Korea
kjw@ysu.ac.kr

Abstract. This paper is a research on various implementations of advanced dynamic signature verification and includes error rates, which are false rejection rate and false acceptance rate, the size of signature verification engine, the size of the characteristic vectors of a signature, the ability to distinguish similar signatures, and so on. We suggest comparison method and the performance results of the signature verification system. We have also implemented web client/server with Java technology, PC (MS Windows), PDA (WinCE) and Smart Phones.

Keywords: Dynamic Signature, Verification, Biometrics, User Authentication, Implementation.

1 Introduction

The ability to identify other individual human beings is fundamental to the security of the family unit. This has been true since the beginning of human history. Members of a tribe needed to identify other members of their tribe quickly, easily, and usually from a distance. They achieved this by using the remembered physical or behavioral characteristics of each tribe member. How a person looked, what they were wearing, how they moved, or combinations of these were used to authenticate the person as a member. The biometric technology [1] allows for a greater reliability of authentication as compared to badges, card readers, or password systems. The chances of an individual losing his biometric information are far less than forgetting a password or losing a card. Through these types of verification, comes an increased role of responsibility, and security.

Dynamic signature verification technology [2, 3, 4, 5] verifies the signer by calculating his writing manner, speed, angle, number of strokes, order, and the down/up movement of pen when the signer inputs his signature with an electronic pen for his authentication.

All biometric techniques have false acceptances generated by the imperfections of the classification method or by errors in the acquisition device. However, dynamic signature verification, using behavioral biometric technique, as compared with physiological biometric techniques such as fingerprint, face, iris or retina, has the additional advantage that a forger with limited information about the true signature cannot deceive the verification algorithm because the multi-dimensional feature

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 203–210.
springerlink.com

information of dynamic signature that is, speed of stroke, size of signature, pressure, variable shape, pen down/up information, and so on decreases the risk of accepting skilled forgeries since such data are not available to the forger.

The rest of this paper is organized as follows: Section 2 describes the dynamic signature verification system; Section 3 describes the comparison method of our system; Section 4 describes performance results of our system: Section 5 describes various implementations for the DSVS; and conclusions follow it in section 6.

2 Dynamic Signature Verification Systems

Dynamic signature verification system (DSVS), like all other biometric verification systems, involves two processing modes: registering and verifying. The registering mode includes three phases: training, testing, and saving. In the training, the user provides signature samples that are used to construct a template (or prototype feature vector) representing some distinctive characteristics of his signature. In the testing, the user provides a new signature to judge authenticity of the presented sample and chooses his own threshold security level.

For the best signature verification, it is important to reduce the range of variation of the true signature and to extend distinctiveness between the true and forgeries.

3 Comparison Method of the Proposed DSVS

Given two signatures to compare, it is natural to ask, "*How similar are they?"* or *"What is their similarity?"* It is intuitive to answer the similarity with a value between 0%-100% and this value should make sense. For example, when we gain the similarity of two signatures as 90%, they should be very close to each other objectively, even though it is subjective to say how similar they are.

No matter what kinds of features are extracted, such a similarity measure is unavoidable. Euclidean distance, DTW (Dynamic Time Warping), or other distances have relative meaning. That is, the distance itself cannot give us any information about similarity without comparing it with other distances.

DTW [7, 8, 9, 10] is one of the best for curve matching with optimal alignment for the dynamic signature verification. Alignment is absolutely necessary, because no user writes exactly the same signature each time. Some differences will always exist in the total length and overall shape.

One of the most important difficulties in authentication using dynamic signatures is the choice of the comparison method. Dynamic signatures are given by a sequence of points sorted with respect to acquisition time. Since two signatures from the same person cannot be completely identical, we must make use of a measure that takes this variability into account. Indeed, two signatures cannot have exactly the same timing. In addition, these timing differences are not linear. Dynamic Time Warping is an interesting tool; it is a method that realizes a point-to-point correspondence. It is

insensitive to small differences in the timing. Calculation distances between signatures with DTW allows one to achieve more flexible, more efficient, and more adaptive verification system than those based on neural networks or Hidden Markov Models [12, 13], as the training phase can be incremental. This aspect is very important when we must enroll our new signature along the years or in a new environment [7].

4 Performance Results

The characteristics of our system are as follows [15 - 18]:

1) Dynamic Time Warping (DTW), well known for excellent pattern matching algorithm, has been modified and applied to this system. Reliability for checking the similarities between signatures is high, and a newly developed, fast algorithm in processing time is adopted in the system. To make access easier, we considered efficient user interface design.
2) Size of feature vector for the signature is very small. It needs 20 bytes-250 bytes of memory capacity to register feature information of a signature on average.
3) Processing time must be fast for verification. In general, with DTW system, it is good to check similarity between patterns, but it has the drawback of increasing processing time because of the complexity of data to be processed. But in our system, we make compressed data and the data structure is well designed, so that it is not affected by time. The verification is processed within 0.001 second with an IBM compatible PC (CPU: 2.0GHz, Main Memory 2GB).
4) Security must be excellent. Through a feedback system of recommendations, the signer can choose among seven security levels, according to skillfulness of the signer.
5) The size of the signature engine is small. Our engine's size is 32KB for Win9x/ME/2000, 6KB for WinCE, and 6KB for JAVA virtual machine. Thus our system can be used in a small, handy device.
6) Especially when using a PDA, Web pad, Tablet PC, Panel PC, smart-phone etc., signature security system is economical and simple because you can install just our software program without purchasing any input devices.
7) Accuracy rate (acceptance rate for true signer and rejection rate for forgery signature) is very high. And error rate is nearly 0 for random forgers.

5 Various Implementations

5.1 Web Client Implementation

We provide two windows (Fig. 1 and Fig. 3) for the dynamic signature verification system. Fig. 1 is a window to save the signer’s signature feature vectors in a remote

database. First step: Signer writes his signature on the white rectangle area and then clicks the 'Register' button. Second step: Signer writes his same signature again and then clicks the 'Test&Verify' button to see recommended security level and degree of similarity in Fig. 2 between the two signatures. With the results of several trials, the signer can choose his security level. Once the signer clicks the 'Save' button, his signature's feature vectors, security level, Resident ID, and password are saved in a remote sign database.

Fig. 1. Interface window for the signature register

Fig. 2. Signature testing window

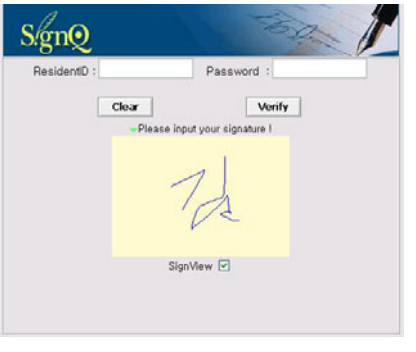


Fig. 3. Interface window for the signature verification

Above Fig. 3 is the user interface window to verify the signer's authentication. The 'SignView' check button is a function to display and erase the writing signature. These interface windows for the DSVS are implemented with JAVA to support various OS platforms.

5.2 Web Server Implementation

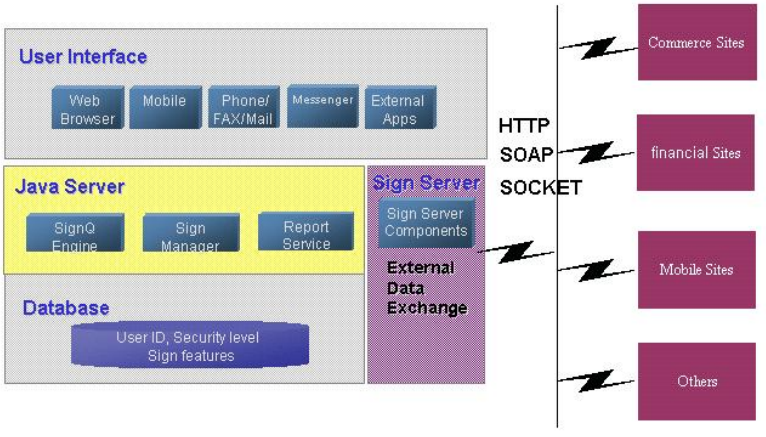


Fig. 4. Components of the sign server and interface structure

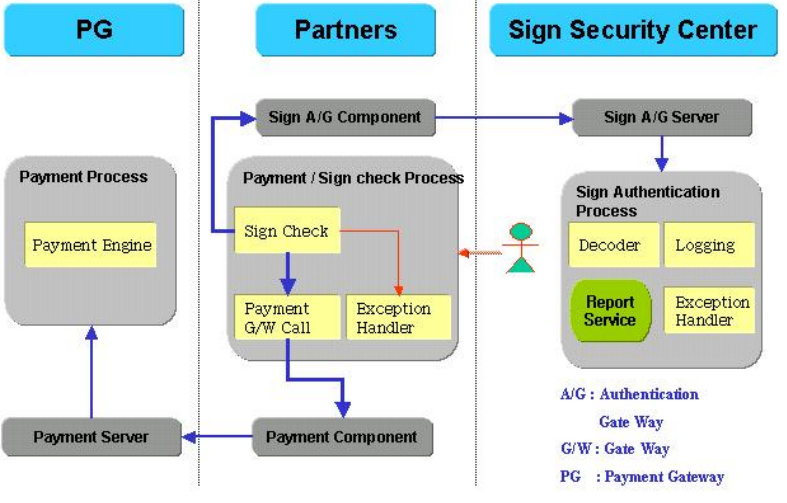


Fig. 5. System architecture of the DSVS

Fig. 4 and Fig. 5 are components of the sign server and interface structure and system architecture of the DSVS respectively.

5.3 Implementation for PDA

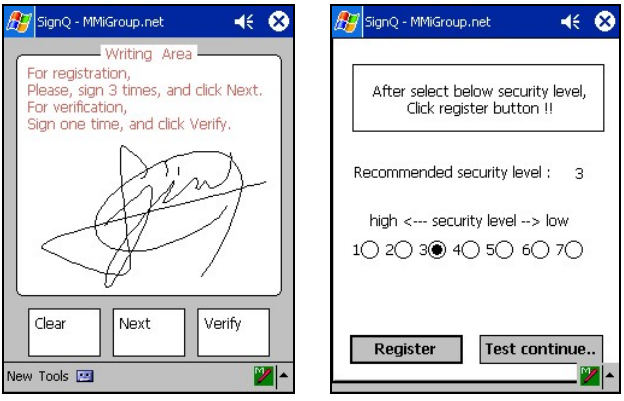


Fig. 6. Interface Design of the DSVS for PDA

Fig. 6 is an implementation and interface design of the DSVS for PDA.

5.4 Implementation for PC Windows

Fig. 7. Interface Design of DSVS for PC Windows

Fig. 7 is an application program to secure files or folders in file system of MS Windows by using the DSVS.

6 Conclusions

We have implemented the DSVS with various Java-based technologies such as Java applet, Java servlet, JSP, HTML, servlet container of Resin, and MySQL database. Also, we have implemented the DSVS as an application program to secure files or folders in PDA, PC Windows and Smart Phones.

The importance of security is emphasized more and more at present, and this system is applicable to the security of a computer, important documents, the access restriction of a network server, on-line shopping, credit cards, military secrets, national administrative security, internet banking, cyber trading, admittance to buildings, personal approval, and so on.

It is quite evident that biometrics is here to stay as the most valuable form of security, which is not only computer-related but also applicable in a plethora of other forms. Markets to be penetrated include using biometrics for passports, birth certificates, forensics, banking, ticketless air travel, computer login, driver's licenses, automobile ignition and unlocking, anti-terrorism, anti-theft, and as a replacement for the archaic PIN and password. As the technologies become increasingly well known, and the market fully embraces these newest forms of biometric security, biometric solutions will inevitably become cheaper, more abundant in the information systems market, and therefore available to almost anybody with a need for enhanced security measures.

Acknowledgements. This thesis was supported by the research funding of Youngsan University.

References

[1] Ruggles, T.: Comparison of Biometric Techniques. Technical Report for The Biometric Consulting Group (1998), http://biometricconsulting.com/bio.htm

[2] Dimauro, G., Impedovo, S., Lucchese, M.G., Modugno, R., Pirlo, G.: Recent Advancements in Automatic Signature Verification. In: Ninth International Workshop on Frontiers in Handwriting Recognition, IWFHR 2004, pp. 179–184 (October 2004)

[3] Plamondon, R., Srihari, S.N.: On-line and Off-line Handwriting Recognition A comprehensive Survey. IEEE Transaction on Pattern Analysis and Machine Intelligence 22(1), 63–78 (2000)

[4] Nalwa, V.S.: Automatic on-line signature verification. Proceedings of the IEEE 85(2), 213–239 (1997)

[5] Jain, A.K., Griess, F.D., Connell, S.D.: Online signature verification. Pattern Recognition 35, 2963–2972 (2002)

[6] Griess, F.D.: Online Signature Verification. Projet Report, Michigan State University, Department of Computer Science and Engineering (2000)

[7] Lei, H., Palla, S., Govindaraju, V.: ER^2: An Intuitive Similarity Measure for On-Line Signature Verification. In: Ninth International Workshop on Frontiers in Handwriting Recognition, IWFHR 2004, pp. 191–195 (2004)

[8] Munich, M.E., Perona, P.: Continuous Dynamic Time Warping for Translation Invariant Curve Alignment with Applications to Signature Verification (1999), http://citeseer.nj.nec.com/munich99continuous.html

[9] Martens, R., Claesen, L.: On-line signature verification by dynamic time-warping. In: The 13th International Conference on Pattern Recognition, pp. 38–42 (1996)

[10] Perizeau, M., Plamondon, R.: A comparative analysis of regional correlation, dynamic time warping and skeletal tree matching for signature verification. IEEE T-PAMI 12(7), 710–717 (1990)

[11] Schimke, S., Vielhauer, C., Dittmann, J.: Using Adapted Levenshtein Distance for On-Line Signature Authentication. In: 17th International Conference on Pattern Recognition, ICPR 2004, vol. 2, pp. 931–934 (August 2004)
[12] Fuentes, M., Garci-Salicetti, S., Dorizzi, B.: On line Signature Verification: Fusion of a Hidden Markov Model and a Neural Network via a Support Machine. In: Proc. of IWFHR-8, Canada, pp. 253–258 (2002)
[13] Dolfing, J.G.A., Aarts, E.H.L., van Oosterhout, J.J.G.M.: Online signature verification with hidden markov models. In: Proceedings of the International Conference on Pattern Recognition, p. 1309 (August 1998)
[14] Ohishi, T., Komiya, Y., Matsumoto, T.: On-line Signature Verification using Pen-Position, Pen-Pressure and Pen Inclination trajectories. In: ICPR 2000, September 03-08, vol. 4 (2000)
[15] Kim, J.W., Cho, H.G., Cha, E.Y.: A Study on the Dynamic Signature Verification System. International Journal of Fuzzy Logic and Intelligent System 4(3), 271–276 (2004)
[16] Kim, J.W., Cho, H.G., Cha, E.Y.: A Study on the Evaluation of Dynamic Signature Verification System. In: ICCSA 2005 Conference Part V. LNCS (2005)
[17] Kim, J.W.: An Advanced Dynamic Signature Verification System for the Latest Smart-Phones. In: Proceedings of International Conference, Mulgrab, Jeju Island, Korea (December 2011)
[18] Kim, J.W., Kim, G.B., Cho, J.H.: A Study on an Advanced Evaluation Method for Dynamic Signature Verification System. International Journal of Maritime Information and Communication Sciences, 140–144 (2010)

Performance of Face Recognition Algorithms on Dummy Faces

Aruni Singh, Shrikant Tiwari, and Sanjay Kumar Singh

Department of Computer Engineering, IT-BHU, Varanasi-India
arunisingh@rocketmail.com,
shrikant.rs.cse@itbhu.ac.in,
sks.cse@itbhu.ac.in

Abstract. Face recognition is becoming increasingly important in the contexts of computer vision, neuroscience, psychology, surveillance, credit card fraud detection, pattern recognition, neural network, content based video processing, assistive devices for visual impaired, etc. Face is a strong biometric trait for identification and hence criminals always try to hide their face by different artificial means such as plastic surgery, disguise and dummy. The availability of a comprehensive face database is crucial to test the performance of these face recognition algorithms. However, while existing publicly-available face databases contain face images with a wide variety of covariates such as poses, illumination, gestures and face occlusions but there is no dummy face database is available in public domain. The contributions of this paper are: i) Preparation of dummy face database of 50 subjects ii) Testing of face recognition algorithms on the dummy face database, iii) Critical analysis of four algorithms on dummy face database.

Keywords: Face recognition, dummy faces, biometrics.

1 Introduction

From very beginning face recognition has become an active area of research in the direction of computer vision, pattern recognition, surveillance, credit card fraud detection, psychology, pattern recognition, neural network, content based video processing, assistive devices for visual impaired etc. Rapid development of face recognition is due to combination of the factors such as development of appropriate algorithms, availability of large facial database and method of evaluating the performance of recognition algorithms [9,7]. Hence Facial Recognition Technology (FRT) has emerged as an attractive solution to address many contemporary requirements for identification [6,16] and verification of identity claims. It brings together the promise of other biometric systems, which attempt to tie identity to individually distinctive features of the body and the more familiar functionality of visual surveillance systems. This paper develops a socio-political analysis that bridges the technical and social-scientific literatures on FRT thus it addresses the unique challenges and concerns that attend its development, evaluation, specific operational uses, contexts, and goals. It highlights the potential and limitations of the technology,

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 211–222.
springerlink.com

noting those tasks for which it seems ready for deployment, those areas where performance obstacles may be overcome by future technological developments and its concern with efficacy extends to ethical considerations [1,7,8]. For the development of FRT face image database is needed. Several researchers have developed so many real face database [10] with a lot of covariates. They have designed and tested many algorithms for recognition and identification of human faces and demonstrated the performance of the algorithms but the performance of face recognition algorithms on dummy and fake faces are not reported in the literature. Since face is prime physiological biometric trait [12] for the identification therefore in the increasing crime in the world, criminals always try to hide their face using fake face, dummy and mask. Hence, the security system will be benefitted.

The main purpose behind spoofing and hiding the original identity by using the masks, disguise or by means of plastic surgery is just to hide the real identity for the purposes of shifting the liability from real to imaginarily face which really does not exist or to adopt the identity of others. This type of situation creates a lot of problems before the courts of law in the administration of criminal justice. Sometimes even such persons (whose mask face has been used by some other person at the time of committing the offence) may be punished who has not committed the offence. Accordingly innocent persons shall be liable for the act of others and thus it will abort the policy or philosophy of criminal justice. This type of spoofing the real face will also attract the amendment of the procedural law and law of evidence. In this paper we have tried to address the performance of face recognition various holistic information based algorithms on dummy or fake faces. The performance evaluation procedures used in this paper will be really encouraging in vitality detection of face image.

This paper has nine sections, section 2 demonstrate the related work and section 3 includes database description consisting preprocessing. Section 4 experimental work and brief description about the algorithms used to identify dummy face. Section 5 contains experimental protocol while section 6 demontrates experimental results and section 7, 8 experimental analysis and future work respectively. Lastly section 9 is conclusion.

2 Related Work

It is found that researchers have worked on face recognition and identification using Principal Component Based techniques [24, 25, 26, 15, 27, 28, 14, 29] and demonstrated the recognition accuracy ranges $\sim 60\,\% - \sim 93\,\%$. Researchers have also worked on FRT using LDA [11, 27, 14, 29] and demonstrated the accuracy ranges $\sim 53\% - \sim 88.75\,\%$, also using ICA [28] and found the accuracy $\sim 73.72\% - \sim 73.72\% - \sim 95.75\%$ and using iSVM [21] demonstrated the accuracy $\sim 86.7 - \sim 100\,\%$ but as to the best of our knowledge the accuracy of face recognition algorithms on fake or dummy face have not ever been demonstrated which is very vulnerable in the area of criminology and fraud detection. To take off this deficiency this work has been incorporated.

3 Database Description

Data acquisition of dummy face is itself a challenging task because unlike real face images we don't have any control over the pose, expression, illumination and occlusion. Thus we have taken the photographs which are available in the public places or market. Further, these images do not follow the standard protocol of face database acquisition. Therefore, our own protocol for data acquisition has been created. We have taken outdoor photographs with 10 Megapixel optical image stabilized Camera, images of 50 subjects have been captured at a distance from nearly 3 feet in an uncontrolled environment. We have captured 10 photographs of each subject from different positions for pose variation with slight variations in illumination due to outdoor snapshot as shown in Fig. 1. Thus the captured images are natural images without imposition of any constraint neither on the targeted subjects nor their surroundings. For database acquisition of dummy faces it took around six months time.

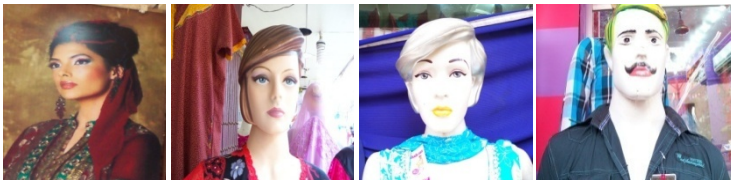


Fig. 1. Original Dummy Faces

3.1 Pre-processing

For the testing of various algorithms requires preprocessing because the photographs of the subjects are taken in uncontrolled environment. We have done following preprocessing steps shown in Fig.2.

We normalized the image up to certain degree so that the face image could be aligned and then cropped out only dummy faces from the dynamic scenes ousting the background. Finally all cropped dummy face image have normalized to set all the subjects at normal gray level illumination and of same size [4].

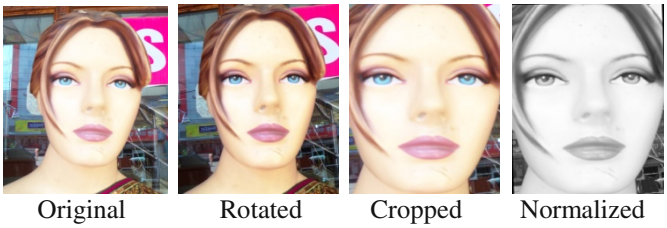


Fig. 2. Preprocessed Images

4 Experimental Work

For the face datasets, we have evaluated the face recognition algorithms PCA, ICA, LDA and iSVM because these algorithms show the holistic performance metrics of accuracy [5]. A brief description of all the four algorithms is given below.

4.1 Principal Component Analysis (PCA)

Principal Component Analysis commonly uses the eigenfaces [13,15] in which the probe and gallery images must be the same size as well as normalized to line up the eyes and mouth of the subjects whining the images. Approach is then used to reduce the dimension of data by the means of image compression basics [17] and provides most effective low dimensional structure of facial pattern. This reduction drops the unuseful information and decomposes the face structure into orthogonal (uncorrelated) components known as eigenfaces. Each face image is represented as weighted sum feature vector of eigenfaces which are stored in 1-D array. A probe image is compared against the gallery image by measuring the distance between their respective feature vectors then matching result has been disclosed. The main advantage of this technique is that it can reduce the data needed to identify the individual to 1/1000th of the data presented [18].

The basis vector are computed from the set of training images I. The average image in I is computed and subtracted from the training images, creating set of data samples

$$i_1, i_2, \ldots\ldots\ldots\ldots i_n \in I - \bar{I} \tag{1}$$

These data samples are arrange in a matrix represented as

$$X = \begin{bmatrix} \begin{bmatrix} \vdots \\ i_1 \\ \vdots \end{bmatrix} & \cdots & \begin{bmatrix} \vdots \\ i_n \\ \vdots \end{bmatrix} \end{bmatrix} \tag{2}$$

XX^T is then the sample covariance matrix for the training images and the principal components of the covariance matrix are computed by solving $R^T(XX^T)R =$ where is the diagonal matrix of eigenvalues and ***R*** is the matrix of orthonormal eigenvectors. Geometrically, ***R*** is a rotation matrix that rotates the original coordinate system onto the eigenvectors, where the eigenvector associated with the largest eigenvalue is the axis of maximum variance, the eigenvector associated with the second largest eigenvalue is the orthogonal axis with the second largest variance, etc. Typically, only the N eigenvectors associated with the largest eigenvalues are used to define the subspace, where N is the desired subspace dimensionality.

Eigenspace terminology, each face image is projected by the top significant eigenvectors to obtain weights which are the best linearly weight the eigenfaces into a representation of the original image. Knowing the weights of the training images and a new test face image, a nearest neighbour approach determines the identity of the face.

4.2 Independent Component Analysis (ICA)

Independent Component Analysis [19] can be viewed as a generalization of PCA [14]. While PCA decorrelates the input data using second-order statistics and thereby generates compressed data with minimum mean-squared reprojection error, ICA minimizes both second-order and higher-order dependencies in the input. It is intimately related to the *blind source separation* (BSS) problem, where the goal is to decompose an observed component into a linear combination of unknown independent components [20, 22]. And then recognition is performed.

4.3 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis is a statistical approach for classifying samples of unknown classes based on training samples with known classes. This technique aims to maximum between-class (across users) variance and minimum within class (within user) variance. In these techniques a block represents a class, and there are a large variations between blocks but little variations within classes.

It searches for those vectors in underlying space that best discriminate among classes (rather than those that best describe the data). More formally given a number of independent features relative to which the data is described. LDA creates a linear combination of these which yields the largest mean difference between desire classes. Mathematically two measures are defined (i) One is called *within-class* scatter matrix which is given by-

$$S_w = \sum_{j=1}^{c} \sum_{i=1}^{N_j} \left(X_i^j - \mu_j\right)\left(X_i^j - \mu_j\right)^T \tag{3}$$

Where X_i^j is the i^{th} sample of class *j*, μ_j is mean of class *j, c* is number of classes, and and $N_j$is number of samples in class *j*, and (ii) Other is called between class scatter matrix

$$S_b = \sum_{j=1}^{c} (\mu_j - \mu)(\mu_j - \mu)^T \tag{4}$$

Where μ represents the mean of all classes. The goal is to maximize the between class measure while minimizing the within class measure. To do this we maximize ratio $\frac{\det |S_b|}{\det |S_w|}$ to prove that if S_w is non-singular matrix then this ration is maximized when the column vectors of the projection matrix, W, are eigenvectors of $S_w^{-1}S_b$. It is noted that - (i) there are at most *c-1* non zero generalised eigenvectors, and so an upper bound of *f* is *c-1* and (ii) require at least *c+t* samples to guarantee that S_w does not become singular. To solve this [23] proposes the use of an intermediate space. In both cases this intermediate space is chosen to be the PCA space. Thus the original t-dimensional space is projected onto an intermediate g-dimension space using PCA and then final f-dimension space LDA.

4.4 Improved Support Vector Machine (ISVM)

Support Vector Machine (SVM) is very popular binary classifier as methods for learning from examples in science and engineering. The performance of SVM is based on the structure of the Riemannian geometry induced by the kernel function. *Amari* in 1999 proposes a method of modifying a Gaussian kernel to improve the performance of a SVM. The idea is to enlarge the spatial resolution around the margin by a conformal mapping, such that the separability between classes is increased [21]. Due to the encouraging results with modifying kernel, this study proposes a novel facial expression recognition approach based on improved SVM (iSVM) by modifying kernels. We have tested this algorithm on our novel dummy database and encouraging result is demonstrated in the figures below.

A nonlinear SVM maps each sample of input space R into a feature space F through a nonlinear mapping φ. The mapping φ defines an embedding of S into F as a curved sub manifold. Denote φ (x) the mapped samples of x in the feature space, a mall vector *dx* is mapped to:

$$\varphi(dx) = \nabla\varphi . dx = \sum_i \frac{\partial}{\partial x^i}\varphi(x)dx(i) \tag{5}$$

The squared length of φ(*dx*) is written as:

$$ds^2 = |\varphi(dx)|^2 = \sum_{i.j} g_{ij}(x)dx^{(i)}dx^{(j)} \tag{6}$$

Where $g_{ij}(x) = \left(\frac{\partial}{\partial x^{(i)}}\varphi(x)\right).\left(\frac{\partial}{\partial x^{(j)}}\varphi(x)\right) = \frac{\partial}{\partial x^{(i)}}.\frac{\partial}{\partial x^{(j)}}.K(x,x')|_{x'=x}$ (7)

In the feature space F, we can increase the margin (or the distances *ds*) between classes to improve the performance of the SVM. Taking the (6) into account, this leads us to increase the Riemannian metric tensor g_{ij} (x) around the boundary and to reduce it around other samples. In view of (7), we can modify the kernel K such that g_{ij} (x) is enlarged around the boundary [21].

5 Experimental Protocol

For our experiment we have taken 10 preprocessed images of each 50 subjects and compressed those images using Gaussian Pyramid [3]. After compression we have prepared the images in the form of Gaussian levels. Level 1 contains compressed images of 100x100 pixels, Level 2 contains images of 50x50 pixels, Level 3 contains images of 25x25 pixels, Level 4 contains images of 13x13 pixels and Level 5 contains images of 7x7 pixels. After compression we have applied the algorithms by including 6 images per subject for training and 4 images per subject for testing.

We have also used both open and closed universe environment for our experiments. In closed universe, every probe images are available in the gallery while in open universe some probe images are not available in the gallery. Both logic [9] reflect very important aspect and report different performance statistics.

6 Experimental Results

For our experiment we have taken 50 subjects and involved 10 photographs of each subject in following scenarios and results are shown in the tables as well as in the figures accordingly. We have taken the result in four scenarios.

(i) For 6 images of each subject as Gallery and 4 images as probe in open universe environment the result or algorithms are shown in Table 1 and Fig. 3.

Table 1. Identification accuracy table in open universe environment

60/40 % Gallery/Probe	Level 1	Level 2	Level 3	Level 4	Level 5
PCA	71.5	72	71	71	51
ICA	70	72.5	70.5	72.5	51
LDA	76.5	73	75	72.5	48.5
iSVM	79	79	79	78.5	63.5

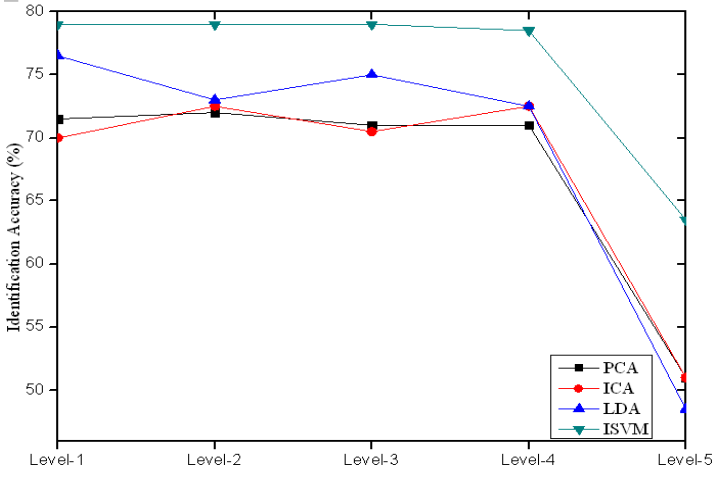


Fig. 3. Identification accuracy graph in open universe environment

(ii) For 8 images of each subject as Gallery and 2 images as probe in open universe environment the results of algorithms are shown in Table 2 and Fig. 4

Table 2. Identification accuracy table in open universe environment

80/20 % Gallery/Probe	Level 1	Level 2	Level 3	Level 4	Level 5
PCA	75	75	79	83	56
ICA	76	76	78	82	55
LDA	77	77	83	82	58
iSVM	84	84	85	83	66

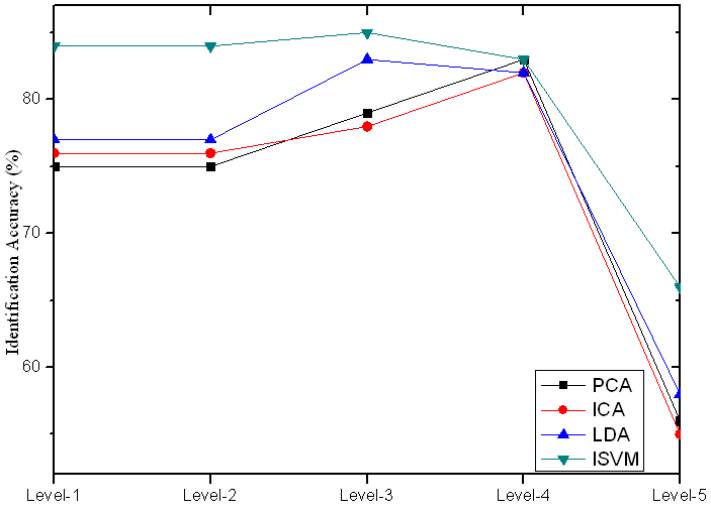


Fig. 4. Identification accuracy graph in open universe environment

(iii) For 6 images of each subject as Gallery and 4 images as probe in close universe environment the results of algorithms are shown in Table 3 and Fig. 5.

Table 3. Identification accuracy in close universe environment

60/40 % Gallery/Probe	Level 1	Level 2	Level 3	Level 4	Level 5
PCA	86.5	86.5	87.5	86.5	76.5
ICA	86.5	87.5	89	85	77
LDA	89.5	89	88	89.5	78.5
iSVM	91	93	92	86.5	79.5

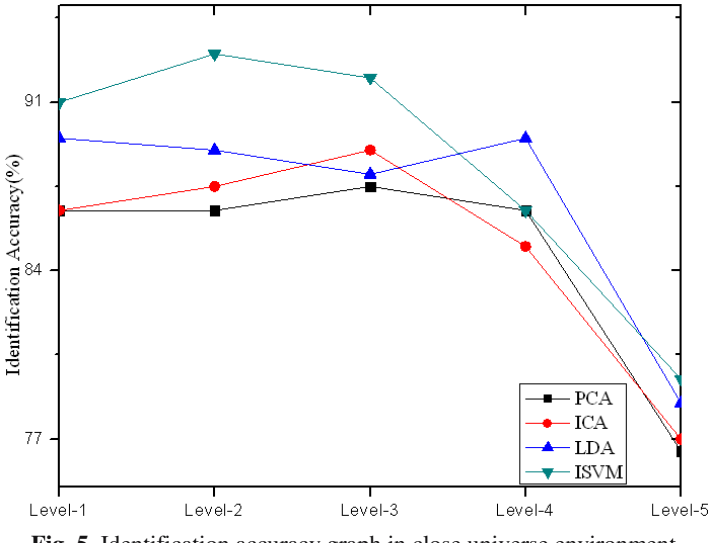


Fig. 5. Identification accuracy graph in close universe environment

(iv) For 8 images of each subject as Gallery and 2 images as probe in close universe environment the results of algorithms are shown in Table 4 and Fig. 6.

Table 4. Identification accuracy table in close universe environment

80/20 % Gallery/Probe	Level 1	Level 2	Level 3	Level 4	Level 5
PCA	90	91	91	88	78
ICA	89	89	91	88	78
LDA	93	92	93	94	84
iSVM	95	95	94	95	82

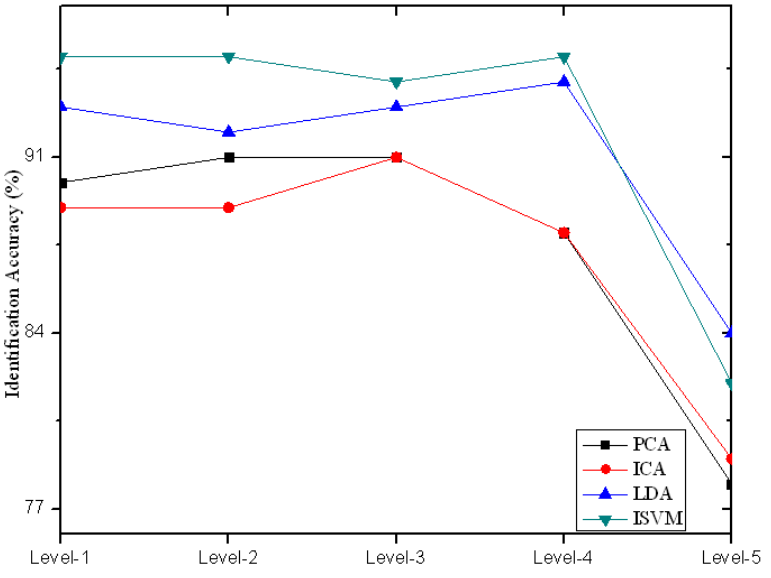


Fig. 6. Identification accuracy graph in close universe environment

The results show that the relative performance of some algorithms is dependent on training conditions (data, protocol) as well as environmental changes. Over the last decade the development of biometric technologies has been greatly promoted by important research techniques.

7 Experimental Analysis

The result shows that the performance varies significantly and iSVM approach has the best performance in level 1 to 4.

- PCA improves the accuracy in with increasing in Gaussian levels because eigenfaces encodes illumination variations.
- LDA is infeasible in large system. In our result the database size in not very large therefore the performance of LDA is in second position after iSVM.

- ICA lacks the performance in these experiments because it works on the basis of independent component. Due to variations in illumination and smoothness in texture database provides less number of independent component which is responsible for degradation of accuracy.
- As we compress the images there is loss of some of its important features and therefore in higher level of compression accuracy decreases.
- When we increase the number of gallery images the algorithms gives the better results.

8 Future Work

The approach described in this paper is initially successful and encouraging in face recognition of dummy faces but more research is to be done in the following domain:

- Size of database is to be increased with illumination variation, pose variation, distance variation, date-variation, expression variation and occlusion variation conditions must be considered while capturing the dummy face of the subjects.
- Our current study reports observed changes due to covariates; however the analysis does not attempt to explain the cause of the effect in detail. Answering the underline cause of the affects will assist in designing more robust face recognition algorithms and then based on their values the most effective algorithm would perform the matching. Alternatively the weighting of an algorithm response would change based on estimated covariates.
- In this respect the evaluation of other types of algorithms are to be done.
- Design and development of new algorithms to distinguish between real and dummy faces.

9 Conclusion

There are so many challenges to develop a comprehensive dummy face database and one of the most fundamental problem in data acquisition is the ability to take consistent, high-quality, repeatable dummy images. In order to compare the performance of some face recognition algorithms on dummy faces we have prepared as well as presented a novel dummy database and tested the matching accuracy of PCA, ICA, LDA and iSVM face recognition algorithms.

The detailed identification results are presented and result demonstrate the factors which affect the identification accuracy are image quality, gallery and probe distribution and uncontrolled image environment. PCA has range of accuracy from (51-72)%, ICA (51-71.50)%, LDA (48.50-76.50)% and iSVM (63.5-79)% at various image compression levels in open universe environment under 60/40 % gallery/probe size. When we increase the gallery size the identification accuracy of each algorithms increases. In this paper, the methodology for creating such database preparation and demonstrate the percentage identification accuracy have been addressed.

References

1. Introna, L.D., Nissenbaum, H.: Facial Recognition Technology. A Servey of Policy and Implementation Issues, CCPR
2. Zhao, W., Chellpa, R., Rosenfield, A., Phillips, P.J.: Face Recognition A Literature Survey
3. Bert, P.J., Adelson, E.H.: The Laplacian Pyramid as Compact Image Code. IEEE Transaction on Communication, COM-31(4) (April 1983)
4. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Pearson Education (2009)
5. Givens, G., Beveridge, J.R., Draper, B.A., Grother, P., Phillips, P.J.: How Features of the Human Face Affect Recognition: A Statistical Comparison of Three Face Recognition Algorithms. In: Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition, vol. 2 (2004)
6. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the Face Recognition Grand Challenge. In: Proc. IEEE Int'l Conf. on Computer Vision and Pattern Recognition, pp. 947–954 (2005)
7. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET Evaluation Methodology for Face Recognition Algorithms. IEEE Transaction on PAMI 22(10), 1090–1104 (2000)
8. Wang, P., Qiang, J., Wayman, J.L.: Modeling and Pridicting face recognition system Performance Based on analysis of similarity score. IEEE Transaction on PAMI 29 (2004)
9. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: Face Evaluation Methodology for Face-Recognition Algorithms. Technical report NISTIR 6264 (January 1999)
10. Dong, H., Gu, N., Pohang: Asian Face Image Database PF01, Intelligent multimedia Lab. Technical Report, San 31, 790-784, Korea
11. Dai, G., Qian, Y.: Face Recognition Using Novel LDA-Based Algorithms
12. Jain, A.K., Hong, L., Pankanti, S.: Biometric Identification. Communication of the ACM 43(2) (February 2000)
13. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs Fisherfaces: class specific linear projection. IEEE Transactions on PAMI 19(7), 711–720 (1997)
14. Martinez, A.M., Kak, A.C.: PCA versus LDA. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(2) (2001)
15. Turk, M., Pentland, A.: Eigenfaces for Recognition. J. Cognitive Neuroscience 3(1) (1991)
16. Samaria, F., Harter, A.: Parameterisation of a Stochastic Model for Human Face Identification. In: Proceedings of 2nd IEEE Workshop on Applications of Computer Vision, Sarasota FL (1994)
17. Sirvoich, L., Kirby, M.: A low dimensional Procedure for Characterization of Human Faces. J. Optical Soc. Am. A 4(3), 519–524 (1987)
18. Cardoso, J.F.: Infomax and Maximum Likelihood for Source Separation. IEEE Letters on Signal Processing 4, 112–114 (1997)
19. Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis. John Wiley & Sons, New York (2001)
20. Hyvärinen, A.: The Fixed-point Algorithm and Maximum Likelihood Estimation for Independent Component Analysis. Neural Processing Letters 10, 1–5 (1999)
21. Liejun, W., Xizhong, Q., Taiyi, Z.: Facial Expression recognition using Support Vector Machine by modifying Kernels. Information Technology Journal 8, 595–599
22. Draper, B.A., Baek, K., Bartlett, M.S., Ross Beveridge, J.R.: Recognizing faces with PCA and ICA. Special issue on face recognition
23. Swets, D.L., Weng, J.J.: Using Discriminant Eigenfaces for Image Retrival. IEEE Transaction on PAMI 18(8), 831–836 (1996)

24. Yambor, W.S.: Analysis of Pca-Based and Fisher Discriminant-Based Image Recognition Algorithms. Technical Report CS-00-103 (July 2000)
25. Ahuja, M.S., Chhabra, S.: Effect of Distance Measures in Pca Based Face, Recognition. International Journal of Enterprise Computing and Business Systems 1(2) (2011) ISSN 2230-8849 (Online)
26. Agarwal, M., Jain, N., Kumar, M., Agrawal, H.: Face Recognition Using Eigen Faces and Artificial Neural Network. International Journal of Computer Theory and Engineering 2(4), 1793–8201 (2010)
27. Dagher, I.: Incremental PCA-LDA Algorithm. International Journal of Biometrics and Bioinformatics (IJBB) 4(2)
28. Draper, B.A., Baek, K., Bartlett, M.S., Ross Beveridge, J.: Recognizing Faces with PCA and ICA
29. Mazanec, J., Melisek, M., Oravec, M., Pavlovicov, J.: Support Vector Machines, Pca and Lda in Face Recognition. Journal of Electrical Engineering 59(4), 203–209 (2008)

Author's Profile

Aruni Singh Asst.Prof. in the Deptt. of Comp.Sc. & Engg., KNIT, Sultanpur, India. Area of interest include biometrics, machine learning, Pattern Classification. Currently pursuing Ph.D. at the Institute of Technology, Banaras Hindu University, Varanasi, India.

Shrikant Tiwari having M.Tech.(Comp. Sc. & Tech.) degree in 2009 from University of Mysore, India. Currently pursuing Ph.D. from IT-BHU, Varanasi, India. His research interests include Biometrics, Image Processing and Pattern Classification.

Dr. Sanjay K. Singh is Associate Professor in Deptt. of Comp. Engg. at IT-BHU, India. He is currently doing research in Biometrics, Pattern Classification, Machine Learning and Image Processing.

Locally Adaptive Regularization for Robust Multiframe Super Resolution Reconstruction

S. Chandra Mohan[1,3], K. Rajan[2], and R. Srinivasan[3]

[1] Dept of IAP, IISc, Bangalore, India
[2] Dept. of Physics, IISc, Bangalore, India
[3] ADE, DRDO, Bangalore, India
rajan@physics.iisc.ernet.in

Abstract. Super resolution reconstruction (SRR) is a post processing technique to correct the degradation of the acquired images due to warping, blur, downsampling and noise. In this paper, image is modeled as Markov random field (MRF) and we propose fuzzy logic filter based on gradient potential (FL) to distinguish between edge and noisy pixels. Based on pixel classification, Tikhonov regularization (TR) or bilateral total variation (BTV) is adopted as a prior in maximum a posteriori (MAP) estimation. Such priors are imperative to obtain a stable solution. Tukey's biweight norm (TBN) is adopted for removing the outliers. The proposed approach is demonstrated on standard test images. Experimental results indicate that the proposed approach performs quite well in terms of visual evaluation and quantitative measurements.

1 Introduction

In imaging applications such as digital photography, video surveillance, remote sensing, military information gathering and medical diagnosis high resolution (HR) image/video is indispensable. In practice, distortions due to optics, imaging sensor and physical constraints such as atmospheric turbulence result low resolution (LR) image/video [1]. A simple solution to increase the spatial resolution of captured images by reducing the pixel size by sensor manufacturing techniques has already been reached a limit. As HR images are important in many fields, computational SRR has emerged as an alternative cost effective approach, which unifies denoising, deblurring and scaling-up tasks. The SRR was first proposed in frequency domain [2]. These frequency domain methods are theoretically simple, computationally efficient and have lessen applications due to their inability to accommodate prior knowledge. To overcome this drawback, many spatial domain approaches [3,4] are being proposed.

Multiframe SRR consists of registration and reconstruction. Available algorithms for registration exhibit various degrees of errors [5]. Recent SRR algorithms focus on robust data fusion, such as L_p norm ($1 \leq p \leq 2$) considering signal independent noise conditions [6,7]. In case of real images with unknown noise models, L_p norm degrades the image quality. To tackle this problem many approaches are proposed [8,9,10,11,12,13]. Many of the existing methods adopt

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 223–234.
springerlink.com

single regularization scheme and regularization parameter (λ) for a complete frame irrespective of region or pixel characteristics. Since most of the natural images contain multiple regions with different spatial characteristics, these algorithms do not provide same performance for all regions. To cater the various regions of the image such as edges/finer details and smooth/flat regions, segmentation [14] or block based methods [15] are proposed to improve the overall performance. However, segmentation methods are complex and block based methods require deblocking mechanism to remove blocking artifacts. In this paper, we propose an alternative approach to distinguish between edge and noisy pixels by utilizing fuzzy logic filter based on gradient potential. Depending on pixel characteristics, Tikhonov regularization or BTV is incorporated as a prior function. Tukey's biweight norm [16,17], which is robust than L_p norm with better outlier rejection capability is employed as data fidelity cost function.

The remainder of the paper is organized as follows. Section 2 describes the forward data model. The proposed pixel classification based on FL is described in section 3 and robust reconstruction approach is illustrated in section 4. Simulations are demonstrated in Section 5. Section 6 concludes this paper.

2 Forward Data Model

The first step in the SRR is to formulate an observation model to replicate the imaging conditions including various degradation factors. This forward model relates the original HR image with the recorded LR frames. The degradation process includes warping (M_k), blurring (B_k), down-sampling (D_k) and AWGN (n_k) terms. The forward model is given as,

$$\begin{aligned} y_1 &= D_1 B_1 M_1 X + n_1 \\ y_2 &= D_2 B_2 M_2 X + n_2 \\ &\vdots \\ y_k &= D_k B_k M_k X + n_k \end{aligned} \tag{1}$$

where, y_k is the acquired LR image, X is the HR image, $k = 1, 2...N$ is number of LR images. Eqn. (1) can be written as

$$\underline{Y_k}^{1} = D_k B_k M_k \underline{X} + \underline{n_k} \tag{2}$$

In digital photography, surveillance imaging applications, B_k combines camera blur (B_k^{cam}), atmospheric turbulence (B_k^{atm}) where B_k^{cam} is dominant. Assuming all the frames are down sampled and blurred by same amount *i.e.*, $\forall D_k = D$, $\forall B_k = B$ and the forward model can be rewritten as,

$$\underline{Y_k} = DBM_k \underline{X} + \underline{n_k} \tag{3}$$

[1] Images are lexicographically ordered.

The problem tackled in this paper is to estimate the HR image $\underline{X}$ from a sequence of LR images $\underline{Y}_k$[2]. In imaging applications, a low cost digital camera (including zoom optics) generates LR images. Here it is assumed that the optical blur functions are already known or estimated. In imaging sensor (CCD), the down-sampling is implemented using averaging strategy. We adopt a more realistic model for M_k, consisting of both translation and rotation. The warp parameters are estimated using Taylor series approximation method [18].

3 Proposed Pixel Classification Based on FL

To reconstruct a better HR image/video, region or pixel characteristic is utmost important. The proposed fuzzy logic based potential approach for image pixel classification consists of fuzzy filtering (FF) and fuzzy smoothing (FS). FF distinguishes intensity variations in the image due to edges and smooth/noisy pixels. It is based on the assumption that a small fuzzy derivative corresponds to a smooth or noisy pixel. On the other hand, the presence of an edge results in a large derivative value [19]. The fuzzy derivative of a pixel at (r,c) along a particular direction is given as,

$$\nabla_F(r,c)\widehat{n} = |y(r,c) - y(\star,\star)|\widehat{n} \tag{4}$$

where, $y(\star,\star)\widehat{n}$ represents nearest pixel value along the unit directional vector $\widehat{n}$. We have considered eight directions and five derivatives along each direction. For a given pixel location at (r,c), to identify an edge along particular direction *i.e.*, $\widehat{n} = \widehat{NE}$ as shown in Fig. 1, the following derivatives are chosen from the set $\mathbf{U}$ and is given as, $\{\nabla(r-2,c-2), \nabla(r-1,c-1), \nabla(r,c), \nabla(r+1,c+1), \nabla(r+2,c+2)\}$. If, majority of the set $\mathbf{U}$ are large and have the same sign,

$$\begin{aligned} &\text{then}, \nabla_F(r,c)\widehat{NE} \ \text{ is large} \\ &\text{else}, \nabla_F(r,c)\widehat{NE} \ \text{ is small} \end{aligned} \tag{5}$$

Here, large derivative is classified as an edge and small derivative corresponds to smooth or noisy pixel. To discriminate between large and small derivative, mean of the absolute value of directional derivative is chosen as a threshold and is given as,

$$\nabla_M = \frac{1}{40}\sum_{\widehat{n}} |\nabla(r,c)\widehat{n}| \tag{6}$$

In this approach, edge and noisy pixels are identified automatically for all images at various noise levels without the knowledge of noise parameters such as standard deviation or variance. This enables us to effectively suppress the noise and preserve the edges during the reconstruction.

[2] Matrix dimensions are listed in [7].

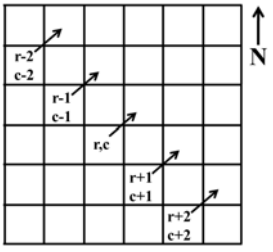


Fig. 1. The directional derivative along $\widehat{NE}$ direction in a 5×5 neighborhood window centered at (r, c)

4 Robust Reconstruction

The statistical problem of estimating unknown $\underline{X}$ is not exclusively based on the information present in $\underline{Y}_k$. It also depends on prior information about the noise, blur kernel and the motion parameters which map the true HR scene to the recorded LR images. It is an inverse problem and finding the solution to the ill-posed and ill-conditioned problem is difficult. A minute change in $\underline{Y}_k$ results in an enormous change in the reconstructed image.

4.1 The Robust Error Norm and Regularization

In SRR, the true HR scene is estimated from the observed data $\underline{Y}_k$. Such inverse problems are ill-posed due to the presence of noise and blurring. Inclusion of prior knowledge to the ill-posed problem improves the reconstruction process and the resulting formulation is the MAP approach. Addition of prior information of the system into the minimization algorithm, guarantees better minimization of the cost functional thereby improving the quality of reconstruction. The MAP estimate of desired HR image is computed as,

$$\underline{\widehat{X}} = \underbrace{\arg \max}_{\underline{X}} \; p(\underline{X}|\underline{Y}_k) \tag{7}$$

By applying Bayes rule,

$$\underline{\widehat{X}} = \underbrace{\arg \max}_{\underline{X}} \frac{p(\underline{Y}_k|\underline{X}) \;\; p(\underline{X})}{p(\underline{Y}_k)} \tag{8}$$

where, $p(\underline{Y}_k)$ is considered as a constant, since the LR images are independent. Using monotonic log function Eqn. (8) can be modified as,

$$\underline{\widehat{X}} = \underbrace{\arg \max}_{\underline{X}} \left\{ \left[\sum_k \log \; p(\underline{Y}_k|\underline{X}) \right] + \log \; p(\underline{X}) \right\} \tag{9}$$

By substituting the likelihood distribution $p(\underline{Y}_k|\underline{X})$ and the prior $p(\underline{X})$ in Gibbs form in Eqn. (9) and maximization of the probability distribution corresponds to the following minimization problem.

$$\widehat{\underline{X}} = \underbrace{\arg \min}_{\underline{X}} \left\{ \|DBM_k\underline{X} - \underline{Y}_k\|_2^2 + \lambda \cdot \mathrm{R}(\underline{X}) \right\} \tag{10}$$

where, the first term is data fidelity and measures the consistency between the estimation and the measurements. The second term is regularization, designed to penalize the solutions that deviate significantly from the estimated HR image. The regularization parameter ($\lambda \geq 0$) balances the contribution of the two terms.

Depending on fuzzy logic based structure analysis, for noisy pixels Tikhonov regularization [20] is applied and is given as,

$$\mathrm{R}_{Tik}(\underline{\mathrm{X}}) = \|\Gamma\underline{X}\|_2^2 \tag{11}$$

where, Γ is a high pass operator such as derivative or Laplacian which forces spatial smoothness of the reconstructed image. For edge pixels, BTV is employed which preserves edges in the reconstructed image and is given as,

$$\mathrm{R}_{BTV}(\underline{\mathrm{X}}) = \underbrace{\sum_{l=-P}^{P}\sum_{m=0}^{P}}_{l+m\geq 0} \alpha^{|m|+|l|}\|\underline{X} - S_x^l S_y^m \underline{X}\|_1 \tag{12}$$

To overcome the drawbacks of L_p norm with better outlier rejection, TBN is adopted [16,17] and is given as,

$$\widehat{\underline{X}} = \underbrace{\arg \min}_{\underline{X}} \left\{ \sum_{k=1}^{N} \rho_{\mathrm{TUKEY}}(DBM_k\underline{X} - \underline{Y}_k) \right\} \tag{13}$$

$$\rho_{\mathrm{TUKEY}}(x) = \begin{cases} \frac{x^2}{T^2} - \frac{x^4}{T^4} + \frac{x^6}{3T^6}, \text{if} |x| \leq T \\ \frac{1}{3}, \text{otherwise} \end{cases} \tag{14}$$

where, T is Tukey's constant parameter which is a soft threshold value. The TBN norm assigns zero or calculated weight to outliers depending on their magnitude. By combining the data fidelity term with regularization, we obtain Eqn. (15).

$$\widehat{\underline{X}} = \underbrace{\arg \min}_{\underline{X}} \left[\sum_{k=1}^{N} \rho_{\mathrm{TUKEY}}(DBM_k\underline{X} - \underline{Y}_k) + \lambda \cdot \right.$$
$$\left. \begin{cases} \|\Gamma\underline{X}\|_2^2, & \text{if}(\nabla_F < \nabla_M) \\ \underbrace{\sum_{l=-P}^{P}\sum_{m=0}^{P}}_{l+m\geq 0} \alpha^{|m|+|l|}\|\underline{X} - S_x^l S_y^m \underline{X}\|_1, & \text{if}(\nabla_F \geq \nabla_M) \end{cases} \right] \tag{15}$$

By steepest descent method, the solution to the cost function is computed by differentiating Eqn. (15) with respect to $\underline{X}$ and HR image is iteratively estimated. Where, α $(0 < \alpha < 1)$ is scalar weight providing spatially decaying effect to the summation of the regularization term, P is size of BTV filter kernel, S_x^l and S_y^m are the operators corresponding to shifting the image $\underline{X}$ by l, m pixels in horizontal, vertical directions respectively. The iteration is terminated when,

$$\frac{\|\widehat{X}_{n+1} - \widehat{X}_n\|^2}{\|\widehat{X}_n\|^2} \leq \varepsilon \tag{16}$$

where, ε is the specified error.

5 Experimental Results

To demonstrate the performance of the proposed approach, we performed experiments on standard monochrome test images. In the first and second experiment, we used cameraman and lena image of size $256 \times 256 \times 1$. All the simulations are carried out using Matlab. To generate the LR images, HR images are warped, blurred by B^{cam}, decimated by a factor of 2, 4[3] in both the directions and AWGS is added to achieve a SNR of 20dB. We assume that, these $\underline{Y}_k$ corresponds to the captured images of our low cost imaging system. In numerical experiments, B^{cam} a Gaussian kernel of size 5×5 with $\sigma = 0.7$, $\alpha = 0.3$, step size $\beta = 0.05$ and $P = 2$ are used. Initially the warp parameters between the LR frames are estimated. The spatial characteristic of the pixels are resolved using the proposed FL approach. TR with large λ is adopted for smooth/noisy pixels to maintain smoothness or to suppress noise. BTV with small λ is bestowed for edge pixels to preserve the finer details and HR image is iteratively reconstructed.

5.1 Discussion

To corroborate the performance of the proposed approach, we compare our results with single frame methods such as bilinear interpolation (BI), spline interpolation (SI), new edge directed interpolation (NEDI) [21] and multi frame methods such as L_2 norm, Tukey error norm with BTV regularization (TTV) [16]. Qualitatively, the proposed approach provides visually pleasing results and are shown in Fig. 2(h), Fig. 3(h). To quantify the reconstructed image quality[4], we have used PSNR[5] and UIQI[6] [22]. Higher values of PSNR and UIQI indicate better reconstruction and are shown in Fig. 6(a), Fig. 6(b). From the experimental results, it is evident that the proposed approach is better in both the subjective and objective measurements. The results demonstrate that the

[3] For D=2, N=6 and D=4, N=24.

[4] Visually appealing result is the basic criterion for parameter selection. Therefore, each experiment is repeated many times and better results are presented.

[5] $\text{PSNR} = 10 \cdot \log_{10}(\frac{255^2}{\text{MSE}})$.

[6] For NEDI, UIQI used the code available at author's site and kindly acknowledged.

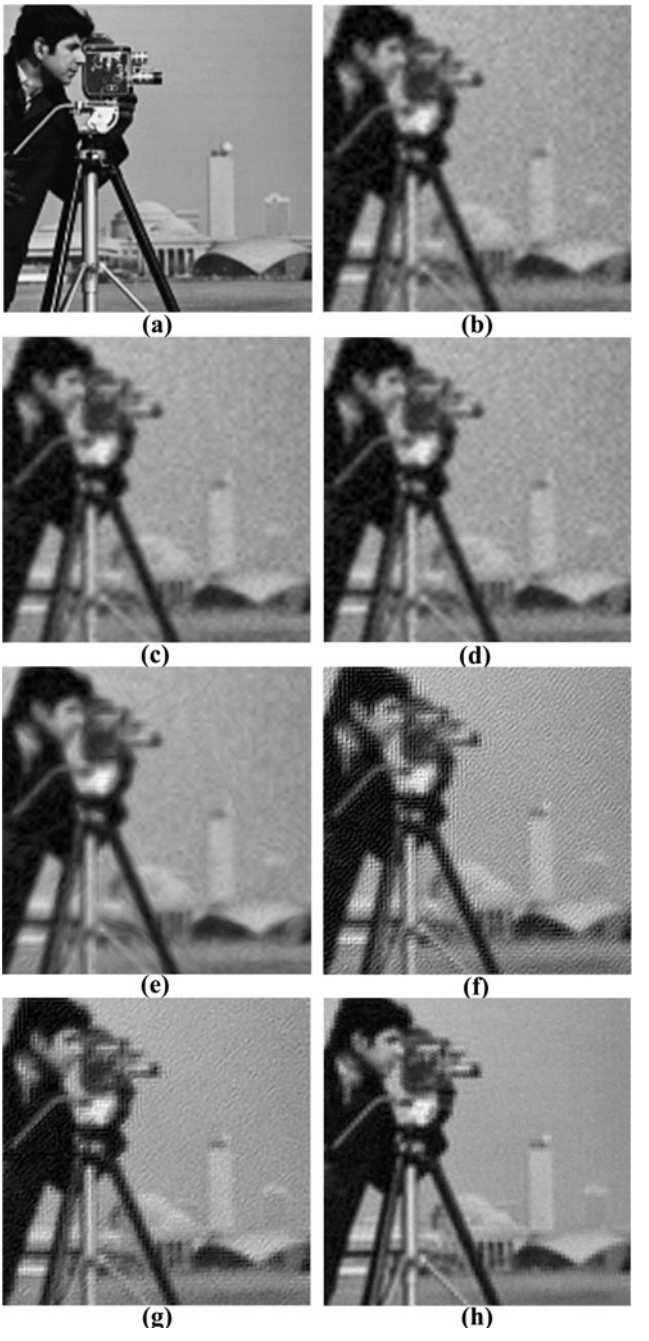


Fig. 2. (a) Region of interest of HR image (size $256 * 256 * 1$), (b) One of LR image (size $128 * 128 * 1$), Reconstructed by (c) Bicubic interpolation, (d) Spline interpolation, (e) NEDI, (f) L_2, (g) TTV, (h) proposed method

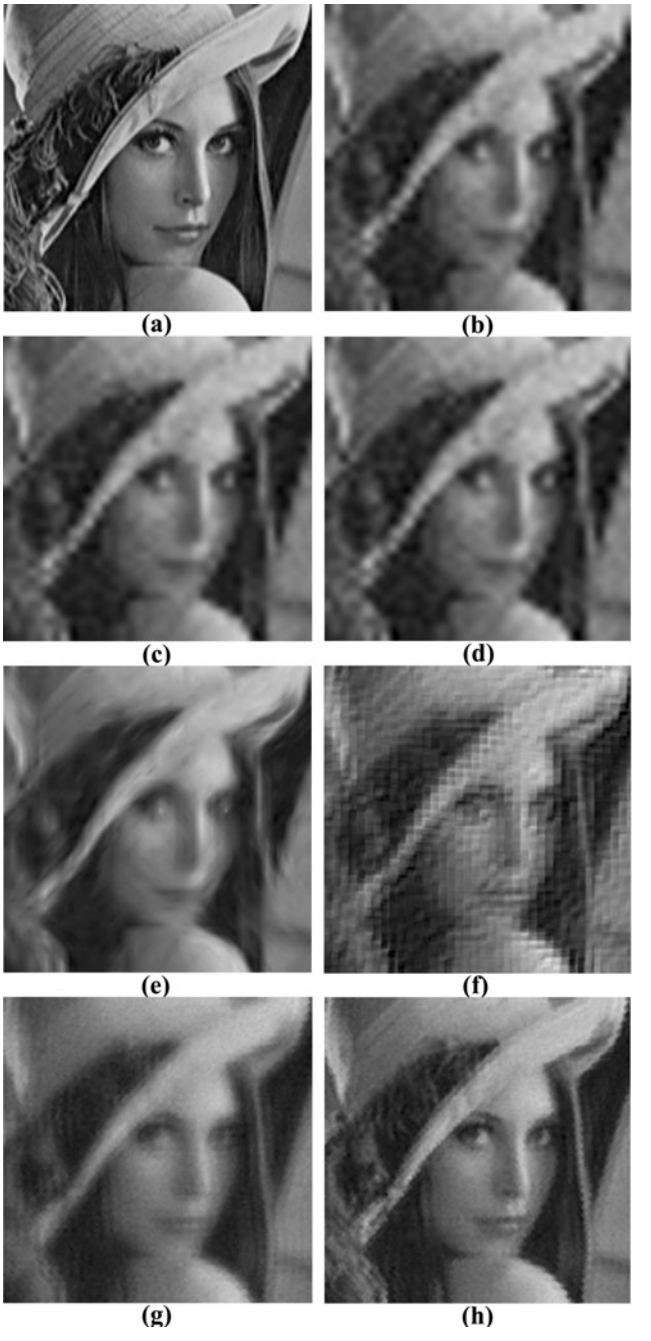


Fig. 3. (a) Region of interest of HR image (size $256 * 256 * 1$), (b) One of LR image (size $64 * 64 * 1$), Reconstructed by (c) Bicubic interpolation, (d) Spline interpolation, (e) NEDI, (f) L_2, (g) TTV, (h) proposed method

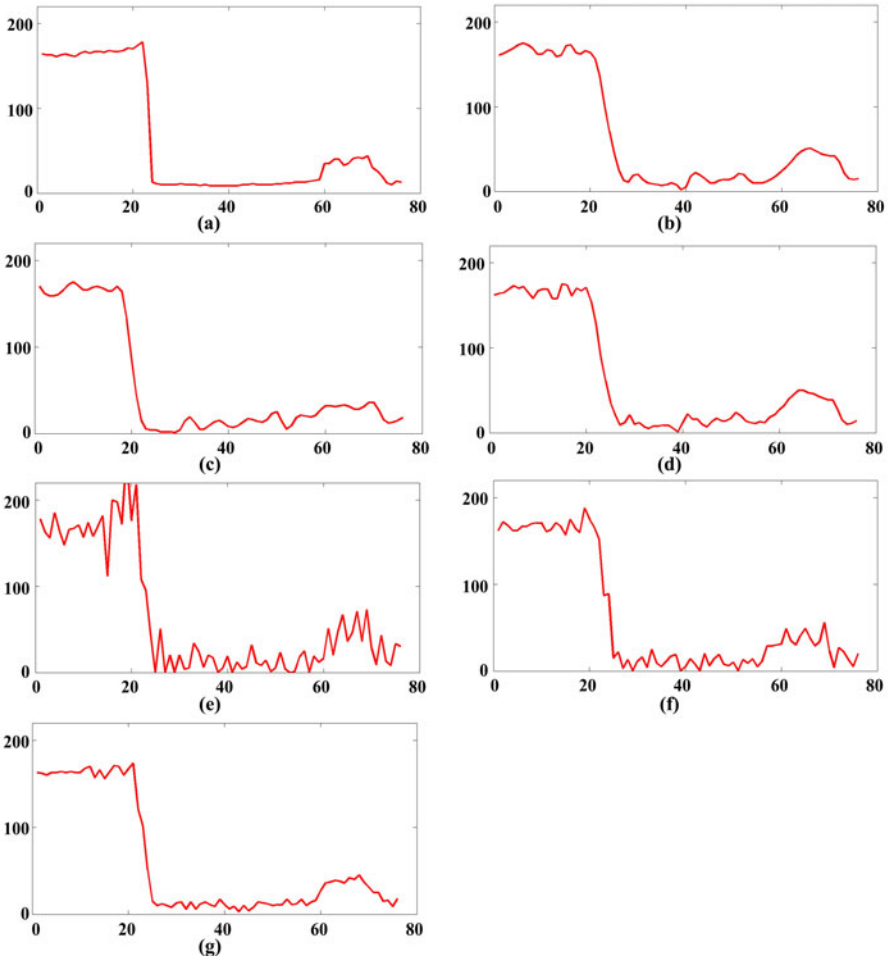


Fig. 4. Line plot (Pixel index vs Intensity) of cameraman image (a) HR image, Reconstructed by (b) Bicubic interpolation, (c) Spline interpolation, (d) NEDI, (e) L_2, (f) TTV, (g) proposed method

proposed framework for SRR using multiple frames preserves the high frequency details, effectively suppresses the noise, rejects outliers, reduces discontinuities and thereby increases the spatial resolution as shown in Fig. 4(g), Fig. 5(g).

5.2 Applicability Issues

The computational complexity of the proposed approach is high. The other major constraint is heuristic estimation of λ, β and T. In our subsequent studies, we focus on methods to reduce the computational cost and estimate the parameters

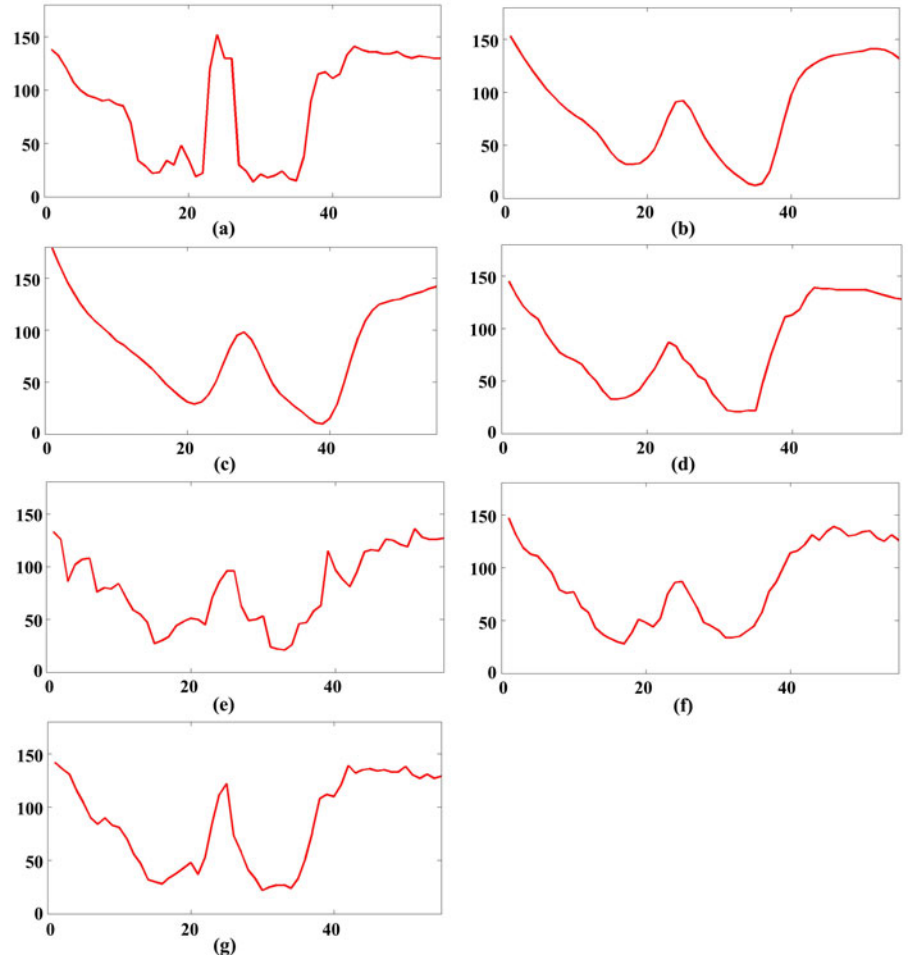


Fig. 5. Line plot (Pixel index vs Intensity) of lena image (a) HR image, Reconstructed by (b) Bicubic interpolation, (c) Spline interpolation, (d) NEDI, (e) L_2, (f) TTV, (g) proposed method

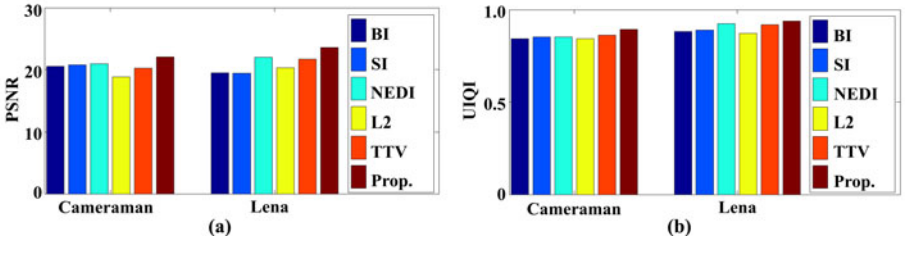


Fig. 6. (a) PSNR (b) UIQI

λ, β and T more systematically. Also we focus on color/biomedical images and implementation in real time by using parallel processing techniques or digital signal processor (DSP).

6 Conclusion

In this paper, a novel locally adaptive SRR approach is proposed for simultaneous noise reduction and preservation of edges/finer details in the reconstructed frames. The spatial characteristic of each pixel is resolved through the fuzzy logic based gradient potential. Regularization for each pixel is adaptively affiliated based on spatial characteristics. Tikhonov regularization with large λ is selected for noisy pixels to suppress noise and BTV with small λ is bestowed for edge pixels to preserve the finer details. The proposed approach gives better results, both quantitatively (higher PSNR, UIQI) and qualitatively. Line plots demonstrate the simultaneous noise reduction and edge preserving capability of the proposed approach. The initial results are encouraging and a lot of work remains to be accomplished. A comprehensive theoretical study for adaptive estimation of numerous parameters in an automated way and real time implementation of the proposed approach will definitely find applications in various fields of science and engineering.

Acknowledgments. The first author acknowledges The Director, ADE, DRDO, Bangalore, India for providing resources to complete this work.

References

1. Bose, N.K., Ng, M.K., Yau, A.C.: A Fast Algorithm for Image Super-Resolution from Blurred Observations. EURASIP Journal on Applied Signal Processing, Article ID 35726, 1–14, (2006), doi: 10.1155/ASP/2006/35726
2. Tsai, R.Y., Huang, T.S.: Multi-frame image restoration and registration. Adv. Computer Vision. Image Process. 1, 317–339 (1984)
3. Tom, B.C., Katsaggelos, A.K.: Reconstruction of a high-resolution image from multiple-degraded misregistered low-resolution images. In: Proc. SPIE, vol. 2308, pp. 971–981 (1994)
4. Schultz, R.R., Stevenson, R.L.: Extraction of high-resolution frames from video sequences. IEEE Trans. Image Processing 5(6), 996–1011 (1996)
5. Zitová, B., Flusser, J.: Image registration methods: a survey. Image and Vision Computing 21(11), 977–1000 (2003)
6. Elad, M., Hel-Or, Y.: A fast super resolution algorithm for pure translational motion and common space invariant blur. IEEE Trans. Image Processing 10(8), 1187–1193 (2001)
7. Farsiu, S., Robinson, D., Elad, M., Milanfar, P.: Fast and robust multi-frame superresolution. IEEE Trans. Image Processing 13(10), 1327–1344 (2004)
8. Hong, M.C., Kang, M.G., Katsaggelos, A.: An iterative weighted regularized algorithm for improving the resolution of video sequences. In: Proc. of Int. Conference Image Processing, ICIP 1997, vol. 2, pp. 474–477 (1997)

9. Hong, M.C., Stathaki, T., Katsaggelos, A.: Iterative regularized least mean mixed norm image restoration. Optical Engineering 41(10), 2515–2524 (2002)
10. He, H., Kondi, L.P.: An image super-resolution algorithm for different error levels per frame. IEEE Trans. Image Processing 15(3), 592–603 (2006)
11. Lee, E.S., Kang, M.G.: Regularized adaptive high-resolution image reconstruction considering inaccurate subpixel registration. IEEE Trans. Image Processing 12(7), 826–837 (2003)
12. Patti, A.J., Sezan, M.I., Tekalp, A.M.: Robust methods for high-quality stills from interlaced video in the presence of dominant motion. IEEE Trans. Circuits Syst. Video Technology 7(2), 328–342 (1997)
13. Hardie, R.C., Barnard, K.J., Bognar, J.G., Armstrong, E., Watson, E.A.: High-resolution image reconstruction from a sequence of rotated and translated frames and its application to an infrared imaging system. Optical Engg. 37(1), 247–260 (1998)
14. Omer, O.A., Tanaka, T.: Region based weighted norm with adaptive regularization for resolution enhancement. Digital Signal Processing 21(4), 508–516 (2011)
15. Zhang, L., Yuan, Q., Shen, H., Li, P.: Multiframe image super resolution adapted with local spatial information. Journal of Optical Society of America A 28, 381–390 (2011)
16. Panagiotopoulou, A., Anastassopoulos, V.: Regularized super resolution image reconstruction employing robust error norms. Optical Engg. 48(11), 117004 (2009)
17. Patanavijit, V., Jitapunkul, S.: A Tukey's Biweight Bayesian Apprach for A Robust Iterative SRR of Image Sequences. In: IEEE Region 10 Conference, TENCON 2007 (2007)
18. Keren, D., Peleg, S., Brada, R.: Image sequence enhancement using sub-pixel displacements. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 1988), vol. CH2605-4, pp. 742–746 (1988)
19. Mondal, P.P., Vicidomini, G., Diaspro, A.: Markov random field aided Bayesian approach for image reconstruction in confocal microscopy. Journal of Applied Physics 102, 044701 (2007)
20. Farsiu, S., Robinson, D., Elad, M., Milanfar, P.: Advances and Challenges in Super-Resolution. Wiley Periodicals, Inc. (2004)
21. Li, X., Orchard, M.T.: New edge directed interpolation. IEEE Trans. Image Processing 10(10), 1521–1527 (2001)
22. Wang, Z., Bovik, A.C.: A Universal Image Quality Index. IEEE Signal Proc. Letters 9(3) (2002)

Improved Watermark Extraction from Audio Signals by Scaling of Internal Noise in DCT Domain

Rajib Kumar Jha[1], Badal Soni[1], Rajlaxmi Chouhan[1], and Kiyoharu Aizawa[2]

[1] Indian Institute of Information Technology, Design & Manufacturing
Dumna Airport Road, PO: Khamariya, Jabalpur 482005, India
jharajib@gmail.com, soni.badal88@gmail.com,
rajlaxmi.chouhan@gmail.com
[2] Department of Information and Communication Engineering,
The Faculty of Engineering, The University of Tokyo, Japan
aizawa@hal.t.u-tokyo.ac.jp

Abstract. Scaling of internal noise in discrete cosine transform (DCT) domain has been presented for copyright protection of audio signals. Watermark as a logo is embedded into the most prominent peaks of the highest energy segment of the audio DCT coefficients. Tuning of the DCT coefficients of the watermarked signal by noise-induced resonance improves the authenticity of the watermarked signal. This scaling is produced by noise-induced resonance, generally known as Dynamic stochastic resonance (DSR). DSR utilizes the noise introduced during different signal processing attacks and it induced here as an iterative process due to which the effect of noise is suppressed and hidden information is enhanced. Response of the proposed extraction scheme suggests increased robustness against various attacks such as noise addition, cropping, re-sampling, re-quantization, MP3 compression, and echo. Comparison with the existing DCT, DWT and SVD techniques shows the better performance in terms of correlation coefficient and visual quality of extracted watermark.

1 Introduction

Recent years have seen that a rapid growth in the accessibility of multimedia content in digital form. The major problem faced by content provider and owners is the protection of their data. Digital watermarking [1] is a growing research field that involves a process to mark digital content by embedding information into the content itself. Watermarking is a process of embedding an information into a digital signal in such a way that it is difficult to remove with the objective of providing authenticity and ownership. A watermark is a signal added to the original digital signal without degrading the quality of original signal, such that it can later be extracted or detected. There are many watermark algorithms in both spatial domain and transform domain [1-6].

A DCT based audio watermarking technique proposed by Dhar *et al.* [2] for copyright protection of audio data by embedding watermark in high energy segments. A blind watermarking scheme through quantization index modulation (QIM) technology

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 235–243.
springerlink.com

was introduced by Zeng *et al.* [3]. A robust audio watermarking in wavelet domain, in SVD domain and time domains respectively are described in [4, 5, 6]. It is more robust to embed a watermark in the transform domain than in time domain due to their decorrelation property. This is why time domain techniques are more susceptible to geometric distortions and various other attacks [4].

Here, dynamic stochastic resonance (DSR) is introduced for extraction of logo from the distorted watermarked audio signal. SR is already used in different fields for different applications [7]. Qinghua *et* al. [8] have used SR phenomenon for line detection from noisy images based on Radon transform. They have shown that the bistable stochastic resonance based Radon transform can easily extract weak lines from very strong noisy images. This SR based Radon transform algorithm is used in the bearing-time record and the LOFAR display. In the same year, Hongler *et al.* [9] reported that the ubiquitous presence of random vibrations in vision systems can be used for edge detection. They showed mathematically and experimentally that the relevant part of the information needed to detect the edges of an image is contained in the modulation of the variance of the output random signal. A constructive action of noise for impulsive noise removal from noisy images is reported by Histace *et al.* [10]. A novel watermarking scheme based on stochastic resonance was reported by Guangchun *et al.* [11]. The watermark is viewed as a weak binary signal, and the median frequency discrete cosine transform (DCT) components of all the 8*X* 8 image blocks are randomly permuted to be an approximate white Gaussian noise (WGN). When the watermark signal which is corrupted by the noise passed through the well-tuned nonlinear system, output signal-to-noise ratio gets improves.

However, the applicability of DSR to audio signal watermarking was not been explored until 2008 by Sun *et al.* [12] in time domain. They have used a parameter Q_{SR} called SR-Degree to obtain the experimental value of the double well parameters which shows the signal as a sub threshold.

In this paper, we have used a concept called dynamic stochastic resonance in which internal noise is scaled to increase the performance of a system. A robust technique for watermark (logo) extraction from audio signals has been proposed in discrete cosine transform domain by scaling of internal noise inherent in the DCT coefficients. In this technique, DCT is applied on the audio signal and watermark is embedded in the selected prominent peaks of highest energy segment of DCT coefficients [2]. Using this technique in the watermark extraction process, the coefficients where watermark was embedded are changed into an enhanced state. An analogy with Benzi's double-well model [13] is used to show the transition of the watermarked coefficients from noisy state to enhanced state.

The prime difference between earlier SR-based audio watermark detection [12] is that, the selection of double-well parameters for noise-induced resonance is done by maximizing the SNR expression of DSR while ensuring the signal is subthreshold mathematically. A gray scale logo extraction from audio signal in DCT domain exploits the property of frequency transformation for watermarking. Embedding and extracting gray-scale image in the audio signal also the difference between the proposed and previous Sun et al. [12].

2 Dynamic Stochastic Resonance

The concept of stochastic resonance was first proposed by Benzi *et al.* 1981 to explain the recurrence of the ice age [13]. The mechanism of stochastic resonance is based on addition of noise. It was traditionally believed that the presence of noise can only make a system worse. However recent studies have shown that in non-linear system noise can induce more ordered regimes and increase the signal-to-noise ratio (SNR), and noise can be used to play a productive role in enhancing the weak input signal. Three components are necessary to induce stochastic resonance, (a) non-linearity (through barrier or threshold), (b) a sub threshold signal, (c) additive noise with a proper *variance.* The SR mechanism shows that at lower noise intensities the weak signal is unable to cross the threshold, thus giving a very low SNR, for large noise intensities the output is dominated by noise, also leading to a low signal to noise ratio. But the moderate noise intensities, the noise allow the signal to cross the threshold giving maximum SNR at some optimum noise level (Fig. 1a).

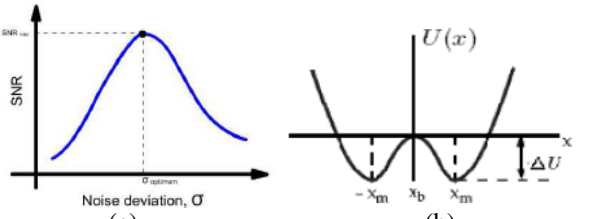


Fig. 1. (a) Signal-to-noise ratio vs. noise standard deviation (b) Bistable double well potential

The bistable-SR model conventionally used by the physicists shall be explored and elaborated in its application to signal detection from watermarked images. The image pixel would transform if mean zero Gaussian fluctuation noise is added; so that the pixel is transferred from weak signal state to enhanced signal state. Such a change of state of pixel under noise can be modeled by Brownian motion of a particle placed in a double well potential system shown in Fig. 1(b).

A classic non-linear dynamic system that exhibits stochastic resonance is modeled with the help of Langevin equation of motion in the form of Eq. 1 given below.

$$\frac{dx(t)}{dt} = -\frac{dU(x)}{dx} + A\sin(\omega t) + \sqrt{P}\ \xi(t) \tag{1}$$

Here P is the noise variance and $\xi(t)$ is the noise, $U(x) = -a\frac{x^2}{2} + b\frac{x^4}{4}$ is a bistable quartic potential as shown in Fig.1b. Here a and b are positive bistable double-well parameters; with the double well minima positions are at $x_m = \pm\sqrt{\frac{a}{b}}$, separated by a barrier of height, $\Delta U = \frac{a^2}{4b}$ when $\xi(t) = 0$. Substituting the value of $U(x)$ in Eq. (1)

$$\frac{dx(t)}{dt} = [ax - bx^3] + A\sin(\omega t) + \sqrt{P}\ \xi(t) \tag{2}$$

Here, *A* and ω are the amplitude and frequency of the periodic signal respectively. It is assumed that signal amplitude is small so that in the absence of noise it is insufficient to force the particle to transit from one well to another, so it will fluctuates around its local minima. A weak periodic forcing and moderate amount of noise is capable of transmitting the particle into another state. Maximum SNR is achieved when $a = 2\sigma_0^2$ the other parameter *b* can be obtained by parameter *a,* for a weak input signal, condition $b < \frac{4a^3}{27}$ is required to ensure sub-threshold condition [13].

By using Eq. (2) and Euler-Maruyama discretized iterative method [14] as follows.

$$x(n+1) = x(n) + \Delta t[\left(ax(n) - bx^3(n)\right) + Input] \tag{3}$$

here Input= *A* sin (ωt) $+\sqrt{P}\,\xi(t)$ can be assumed to be the noisy watermarked coefficient as they contain signal as well as noise [14]. This iterative tuning processing when applied to the DCT coefficients of a watermarked audio signal causes scaling of internal noise which is inherent in the DCT coefficients.

3 Performance Characteristic

Performance of audio watermarking algorithms is commonly evaluated with respect to two common metrics: imperceptibility (inaudibility) and robustness [4].

Imperceptibility (Inaudibility) is related to the perceptual quality of the embedded watermark data within the original audio signal. It ensures that the quality of the signal is not perceivably distorted and the watermark is imperceptible to a listener. To measure imperceptibility, Signal-to-Noise Ratio (*SNR*) is computed between cover signal and watermarked signal to be used as an objective measure. According to [2], for good imperceptibility, *SNR* should be in range 13dB to 24dB.

Robustness: To gauge the quality of extraction, correlation coefficient ρ, between original watermark and extracted watermark is computed. ρ take values between 0 (random relationship) to 1 (perfect match).

4 Watermark Embedding Algorithm

The steps of watermark embedding are as proposed by [2]. The input watermark is a gray scale logo of size (66×66). To embed this image in audio signal, it is reshaped into an array of dimension (4356×1). In the proposed technique, watermark is embed into most prominent peaks of highest energy segment because most of the common signal processing attacks affect low energy segment and this is why the highest energy segment can be considered to be robust to such attacks. The steps of embedding are:

Step 1. Apply DCT to the original audio signal. These DCT coefficients are segmented into arbitrary number of segments and energy of each segment is calculated by $EG = \sum_i |U(i)^2|$. Here *EG* is the total energy of U^{th} segment. Find the highest energy segment and the most prominent peaks in the highest energy segment.

Step 2. The watermark is inserted or embedded into the selected *n* most prominent peaks of highest energy segment, where *n* the length of watermark. The watermarks are embedded into the highest *n* DCT coefficient by the following equation

$$U_{i=}' U_i(1 + \alpha W_i) \tag{4}$$

Here W_i is the watermark, U_i is the magnitude coefficient into which the watermark is to be embedded, U_i' is the adjusted magnitude coefficient, α is the watermark amplification factor, taken as 0.30.

Step 3. Inverse DCT is applied to the modified (watermarked) coefficients to obtain the watermarked audio signal.

5 Proposed Watermark Extraction Algorithm

SR is produced by using the degradation added during various signal processing attacks. The steps of watermark extraction are as follows.

Step 1. Attacked watermarked signal is transform into discrete cosine transform domain.
Step 2. Find all those most prominent peaks in the DCT coefficient which were used in watermarked embedding process as U_i'.
Step 3. Now scale the selected DCT coefficients by applying the DSR iterative equation (3).

Here *Input* is the DCT coefficients, where watermark was embedded. From mathematical simplicity, iteration starts with initial condition $x(0) = 0$ and parameter $a = 2\sigma_0^2$ is taken following the condition of maximizing of SNR [14]. Here σ_0^2 is the variance of U_i'. Using iterative equation given in Eq. (3) calculate the tuned most prominent DCT coefficients. Here $x(n+1)$ *i*s the DSR-tuned, in other words, scaled set of coefficients where *n* is the number of iterations. The experimentally value of the bistable parameters are b=0.001×4a^3/27 and $\Delta t = 0.03$.

Step 4. Watermark coefficients are extracted from these DSR-tuned coefficients using inverse operation of watermark embedding process.

$$w_i^* = \left(\frac{x(n+1)}{U_i} - 1\right)/\alpha \tag{5}$$

Here $x(n+1)$ and U_i are the watermarked DSR based tuned coefficient and original signal coefficient respectively; Watermark sequence is then generated $w^* = \{w_1^*, w_2^*, w_3^*, \ldots \ldots w_n^*\}$ after every iteration.

Step 5. Every iteration reconstructs the gray-scale image watermark. Correlation coefficient like ρ is computed between original watermark and the extracted watermark. To make the algorithm adaptive, the iterative process is continued as long as ρ keeps increasing and stops for that values of *n* (iteration*)* where it reaches a peak and

starts decreasing henceforth. This value of iteration is taken as the optimum number of iteration, n_{opt}.

6 Experimental Results

Each audio signal ((a) classical music (b) human voice (c) rock music) contains 70800 samples with duration of 12 seconds. A 66 × 66 grayscale image has been used as the watermark. The original watermark image is shown in Fig. 2d. The watermarked signals of Fig. 2a-2c have been shown in Fig. 2e-2g along with their *SNR* values. It is apparent from that all *SNR* that watermark was found to be nearly imperceptible. This was in agreement with observations made by objective listening. Extracted watermark from classical music signal in the absence of any attacks has been shown in Fig. 2h. Graph of correlation coefficient (ρ) as a function of iteration count have been shown in Fig. 4.

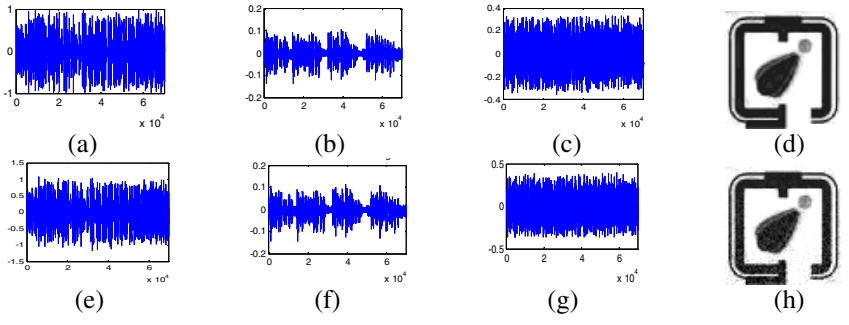


Fig. 2. Cover audio signal (a) Classical music (b) Human Voice (c) Rock music (d) Original watermark. Watermarked audio signals for α = 0.3 (e) Classical music, SNR =17.47 dB (f) Human Voice, SNR=20.49 dB (g) Rock music, SNR = 18.85 dB (h) Extracted watermark from (e) without attack (**ρ** = 0.9998)

Table 1. Correlation coefficient of extracted watermark using proposed watermark extraction technique for three different cover audio signals

Attacks	**Correlation coefficient (ρ)**		
	Classical music	Human voice	Rock music
Noise addition (30 dB)	**0.9801**	0.9872	0.9257
Echo (20%, 100ms)	**0.9901**	0.9774	0.9892
Cropping (10 %)	**0.9777**	0.9770	0.9860
Re-sampling (50%)	**0.9525**	0.9245	0.9552
MP3 Compression	**0.9693**	0.9367	0.9651
Zero-crossing (thresh=0.9)	**0.9843**	0.9803	0.9654
Median filter size (3×3)	**0.7854**	0.7309	0.6545
Low pass filter (Butterworth order 5)	**0.9887**	0.9907	0.9713

Table 2. Correlation coefficient of extracted watermark using proposed watermark extraction technique and existing methods for different attacks (on classical music)

Types of Attacks	Correlation coefficient (ρ)			
	DSR-DCT	DCT [2]	DWT [4]	SVD [5]
Noise addition	**0.9801**	0.9113	0.9554	0.9203
Echo	**0.9901**	0.9462	0.9040	0.8295
Cropping	**0.9777**	0.8939	0.8947	0.8740
Re-sampling	**0.9525**	0.8137	0.8567	0.8455
Compression	**0.9693**	0.7394	0.8399	0.7621
Zero- crossing	**0.9843**	0.8530	0.9308	0.8513
Median filter	**0.7854**	0.3607	0.4094	0.4521
Low pass filter	**0.9887**	0.9619	0.9376	0.8973

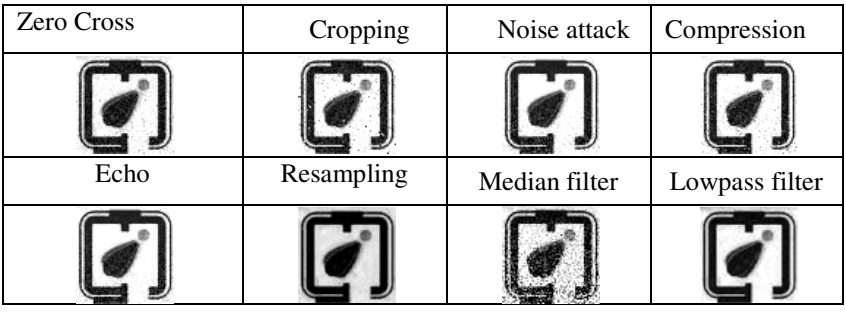


Fig. 3. Extracted watermarks using DSR-DCT on Classical music audio signal

7 Discussion

Salient features of the proposed technique have been discussed in this section.

7.1 Quality of Extracted Watermark

The proposed DSR-based watermark extraction technique uses internal noise of the attacked signal and gives good robustness quality against various attacks [11], Table 1 shows correlation coefficient values obtained for three cover audio signals, *Classical music*, *Human voice* and *Rock music* (as shown in Fig. 2e-2g) in the presence of various attacks. When the watermarked signal is subjected to attacks like gaussian noise, echo, cropping, re-sampling, MP3 compression, zero-crossing, low pass and median filtering, the proposed technique is observed to reach very high correlation shown as high as 0.9901 (for echo attack). Table 2 shows correlation coefficient values obtained using the proposed technique for various attacks in comparison with existing watermark extraction techniques in DCT, DWT and SVD domains respectively [2, 4, 5]. In comparison with existing technique the proposed technique is found to give better performance for almost all attacks in terms of correlation coefficient and visual quality of extracted watermarks. Fig. 4 show that the maximum value of ρ (for additive white gaussian noise attack) is achieved at n_{opt}=18.

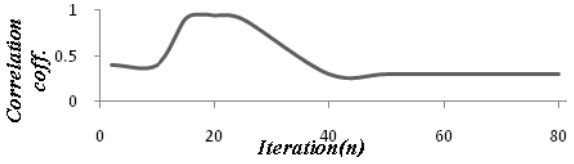


Fig. 4. Graph between correlation coefficient (ρ) wrt iteration (n)

7.2 Mechanism of DSR

The basic increase in robustness is attributed to the nature in which tuning/scaling affects the distribution of DCT coefficients on which watermark was embedded. It can be inferred from Fig. 5, that attack (here, additive white Gaussian noise 30 dB) flattens the distribution of DCT coefficients of the watermarked signal. Scaling of the attacked DCT coefficients (or the internal noise inherent in them) using the iterative equation increases the spread of distribution. This causes an increase in the energy of hidden data with successive iterations enabling easy extraction of watermark.

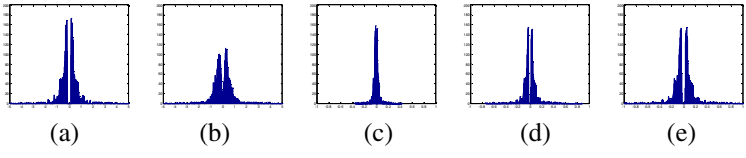


Fig. 5. (a) DCT coefficient distribution of (a) watermarked cover signal (b) after white Gaussian noise attack (c) after DSR iteration n=7 (d) after DSR iteration n=13 (e) after DSR iteration n=18

8 Conclusions

A robust technique for logo extraction from audio signals in DCT domain by scaling of internal noise was proposed and investigated in this paper. Watermark embedding in high energy regions of the audio signal defends the inaudibility of the watermark in the watermarked audio signal. Use of dynamic stochastic resonance in the extraction process to scale the internal noise due to various degradations suggested a drastic improvement in robustness of the extraction process. The iterative equation (analogous to motion of a particle in double-well) was found to scale or tune the degradation or noise introduced during various attacks by producing a noise-induced transition from the poor state of coefficients to enhanced state. Experimental results show that our proposed technique achieves correlation coefficient values larger than those obtained using existing DCT, DWT and SVD-based techniques for almost all types of attacks. An adaptive DSR-iterative procedure gives remarkable performance giving minimum computational complexity in terms of iteration count and can be considered suitable for robust logo extraction from any audio signal.

References

1. Tewfik, H.: Digital watermarking. IEEE Signal Processing Magazine 17, 17–88 (2005)
2. Dhar, P.K., Khan, M.I., Ahmad, S.: A new DCT-based watermarking method for copyright protection of digital audio. International Journal of Computer Science and Information Technology 2(5), 91–101 (2010)
3. Zeng, G., Qiu, Z.: Audio watermarking in DCT-Embedding Strategy and Algorithm. In: Proc. IEEE International Conference on Signal Processing, pp. 2193–2196 (2008)
4. Al-Hai, A., Mohammad, A., Bata, L.: DWT based audio watermarking. International Arab Journal of Information Technology 8(3), 326–333 (2011)
5. Ozer, H., Sankur, B.: An SVD-Based Audio Watermarking Technique. In: IEEE Signal Proc-essing and Communications Applications Conference, pp. 452–455 (2005)
6. Bassia, P., Pitas, I., Nikolaidis, N.: Robust audio watermarking in the time Domain. IEEE Transactions on Multimedia 2(3), 232–241 (2001)
7. Gammaitoni, L., Hanggi, P., Jung, P., Marchesoni, F.: Stochastic resonance. Review of Modern Physics 70, 223–287 (1998)
8. Ye, Q., Huang, H., He, X., Zhang, C.: A SR-based Radon transform to extract weak lines from noise images. In: Proc. of IEEE ICIP, vol. 5, pp. 1849–1852 (2003)
9. Hongler, M., Meneses, Y., Beyeler, A., Jacot, J.: Resonant retina: Exploiting vibration noise to optimally detect edges in an image. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(9), 1051–1062 (2003)
10. Histace, A., Rousseau, D.: Constructive action of noise for impulsive noise removal in scalar images. Electronics Letters 42(7), 1–2 (2006)
11. Wu, G., Qiu, Z.: A novel watermarking scheme based on stochastic resonance. In: IEEE 8th International Conference on Signal Processing, vol. 2, pp. 1–4 (2006)
12. Sun, S., Ming, J., Xing-qiong, L.: A blind audio watermarking based on stochastic resonance signal processor. In: Proc. IEEE International Conference on Signal Processing, pp. 2185–2188 (2008)
13. Benzi, R., Sutera, A., Vulpiani, A.: The mechanism of stochastic resonance. Journal of Physics A 14, L453–L457 (1981)
14. Chouhan, R., Jha, R.K., Chaturvedi, A., Yamasaki, T., Aizawa, K.: Robust Watermark Extraction Using SVD-Based Dynamic Stochastic Resonance. In: Proc. IEEE International Conference on Image Processing, vol. 11, pp. 2801–2804 (2011)

Performance of Adders with Logical Optimization in FPGA

R. Uma and P. Dhavachelvan

Department of Computer Science, School of Engineering, Pondicherry University, Puducherry, India
uma.ramadass1@gmail.com, dhavachelvan@gmail.com

Abstract. Addition is an indispensable operation for any high speed digital system, digital signal processing or control system. Therefore, careful optimization of the adder is of the greatest importance. This optimization can be attained in two levels; it can be circuit or logic optimization. In circuit optimization the size of transistors are manipulated, where as in logic optimization the Boolean equations are rearranged (or manipulated) to optimize speed, area and power consumption. In this work technology independent logic optimization is used to design 1-bit full adder with 20 different Boolean expressions and its performance is analyzed in terms of transistor count, delay and power dissipation using Tanner EDA with TSMC MOSIS 250nm technology. All the Boolean expressions are realized in terms of CMOS logic. From the analysis the optimized equation is chosen to construct a full adder circuit in terms of multiplexer. This logic optimized multiplexer based adders are incorporated in selected existing adders like ripple carry adder, carry look-ahead adder, carry skip adder, carry select adder, carry increment adder and carry save adder and its performance is analyzed in terms of area (slices used) and maximum combinational path delay as a function of size. The target FPGA device chosen for the implementation of these adders was Xilinx ISE 12.1 Spartan3E XC3S500-5FG320. Each adder type was implemented with bit sizes of: 8, 16, 32, 64 bits. This variety of sizes will provide with more insight about the performance of each adder in terms of area and delay as a function of size.

Keywords: Digital signal processing, Carry ripple adder, Carry Look-ahead adder, FPGA, VLSI, logic optimization.

1 Introduction

In most of the digital systems, adders are the fundamental component in the design of application specific integrated circuits like RISC processors, digital signal processors (DSP), microprocessors etc. The design criterion of a full adder cell is usually multifold. Transistor count is, of course, a primary concern which largely affects the design complexity of many function units such as multiplier and Arithmetic logic unit (ALU). The basic principle in designing digital adder circuit hovers around reducing the required hardware thus reducing the cost too. To achieve this, logical optimization helps to obtaining minimum number of literals to minimizing the transistor count and the power consumption and increasing the speed of operation.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 245–254.
springerlink.com

A logic expression can be expressed in various logic forms, which differ in literal counts. In widely-used MOS circuits, the number of transistors to implement a Boolean expression is directly proportional to literal counts in its logic form[1-2]. Thus, a logic optimization is simply to derive a logic form with the fewest literals. Logic level optimization is the design task where an RTL circuit description is optimized in terms of area, delay and power. Conventionally, a logic level optimization can be achieved in two steps; they are Technology Independent (TI) and Technology – Dependent (TD) optimization. In the former method the circuit's Boolean description is optimized ignoring the technology in which the circuit will be implemented. In the second method, the output of the technology independent optimization step (i.e. optimized Boolean network) is optimized considering the adopted technology. During the TI step there is much flexibility to restructure circuit logic to minimize the number of nodes and literals, thereby reducing the area of the circuit. During this stage the circuit can be most effectively restructured to meet the specified delay constraints critical for circuit performance. During the TD step, the delay characteristics of the target library are available, but very few restructuring of the circuit is possible.

In this paper, we proposed 20 different Boolean expressions (logic construction) to implement a 1-bit full adder circuit. All the Boolean expressions are realized in terms of CMOS logic. The optimization method used in this work is technology independent optimization step. These Boolean logic realization and performances are analyzed in terms of transistor count, delay and power dissipation using Tanner EDA with TSMC MOSIS 250nm technology. From this analysis the optimized equation is selected and it is implemented in terms of multiplexers and it is incorporated in selected existing adder topologies like ripple carry adder, carry look-ahead adder, carry skip adder, carry select adder, carry increment adder and carry save adder and its performance is analyzed in terms of area (slices used) and maximum combinational path delay as a function of size. Performance comparison of existing and logic optimized schemes are analyzed on cell-based VLSI technologies, such as standard-cell based FPGAs. The cell-based approach is justified by its wide-spread use in the ASIC design community and its compatibility with hardware synthesis, which in turns satisfies the demand for ever higher productivity. This work presents the significance of adder comparison in terms of CLBs occupied and its maximum combinational delay exist in adder topology.

The organization of the paper is as follows: The section 2, describes the existing adder topologies. Section 3, presents the mathematical Boolean expression for the design of 1-bit full adder cell. Section 4 presents the simulation and analysis of full adder using Tanner EDA. Section 5 presents the FPGA implementation of different adder topologies. Section 6 gives the summary of comparison. Finally the conclusion is presented in section 7.

2 Review of Existing Adder Topology

Most of the VLSI applications, such as digital signal processing, image and video processing, and microprocessors, extensively use arithmetic operations. Addition, subtraction, multiplication, and multiply and accumulate (MAC) are examples of the most commonly used operations. The 1-bit full-adder cell is the building block of all these

modules. Thus, enhancing its performance is critical for enhancing the overall module performance. This section presents the overview of the existing adder topologies.

The adder topology is present in literature [3-12], Ripple Carry Adder (RCA) is the simplest, but slowest adders with $O(n)$ area and $O(n)$ delay, where n is the operand size in bits. Carry Look-Ahead (CLA) have $O(n\log(n))$ area and $O(\log(n))$ delay, but typically suffer from irregular layout. On the other hand, Carry Skip Adder, carry increment and carry select have $O(n)$ area and $O(n^{l+2/l+1})$ delay provides a good compromise in terms of area and delay, along with a simple and regular layout. Carry save adder have O(n) area and $O(\log n)$ delay. The ripple carry adder, the most basic of flavours, is at the one extreme of the spectrum with the least amount of CLBs but the highest delay. CLA adders can be realized in two gate levels provided there is no limit on fan in/out. The carry select adders reduce the computation time by pre-computing the sum for all possible carry bit values (ie '0' and '1'). After the carry becomes available the correct sum is selected using multiplexer. Carry select adders are in the class of fast adders, but they suffer from fan-out limitation since the number of multiplexers that need to be driven by the carry signal increases exponentially. In the worst case, a carry signal is used to select n/2 multiplexers in an n-bit adder. When three or more operands are to be added simultaneously using two operand adders, the time consuming carry propagation must be repeated several times. If the number of operands is 'k', then carries have to propagate (k-1) times.

3 Mathematical Equations for Full Adder

A full adder is a combinational circuit that performs the arithmetic sum of three bits: A, B and a carry in, C, from a previous addition produces the corresponding SUM, S, and a carry out, CARRY. The various equations for SUM and CARRY are given below

$$SUM = A \oplus B \oplus C \quad (1)$$

$$SUM = \overline{A \oplus B \oplus C} \quad (2)$$

$$SUM = \left(\overline{A}B + \overline{B}A\right)C + \left(\overline{A}B + A\overline{B}\right)\overline{C} \quad (3)$$

$$SUM = \left(\overline{\overline{A}\,\overline{B} + AB}\right)\overline{C} + \left(\overline{A}\,\overline{B} + AB\right)C \quad (4)$$

$$CARRY = (A \oplus B)C + AB \quad (5)$$

$$CARRY = \left(\overline{A \oplus B}\right)C + AB \quad (6)$$

$$CARRY = \overline{\left(\overline{AB} + \overline{(A \oplus B)}\,\overline{C}\right)} \quad (7)$$

$$CARRY = \overline{\overline{AB} \bullet \overline{AC} \bullet \overline{BC}} \quad (8)$$

$$CARRY = \overline{A \oplus B} \bullet B + A \oplus B \bullet C \quad (9)$$

$$CARRY = \overline{(A \oplus B) \bullet C \oplus \overline{AB}} \quad (10)$$

$$CARRY = C \bullet (A \bullet B) + C \bullet (A + B) \quad (11)$$

$$CARRY = AB + AC + BC \quad (12)$$

$$CARRY = \overline{\overline{(A \oplus B)C} \bullet \overline{AB}} \quad (13)$$

$$CARRY = AB + \overline{A \oplus B} \bullet C \quad (14)$$

$$CARRY = \overline{AB \bullet AC \bullet BC} \quad (15)$$

$$CARRY = \overline{\overline{A \oplus B} \bullet \overline{AC} \bullet \overline{BC}} \quad (16)$$

$$CARRY = \overline{\overline{AB} \bullet \overline{\overline{A} + \overline{B} \bullet C}} \quad (17)$$

$$CARRY = \overline{\overline{A \oplus B \bullet C} \oplus \overline{AB}} \quad (18)$$

$$CARRY = \overline{\overline{A} + \overline{B}} + \overline{\overline{A \oplus B} + \overline{C}} \quad (19)$$

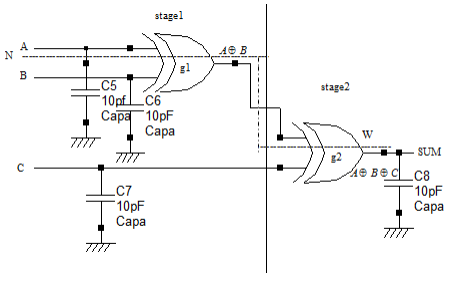
$$CARRY = \overline{\overline{\left(\overline{A \oplus B}\right)C} \oplus \overline{AB}} \quad (20)$$

In this work 20 different Boolean expressions are formulated. Using this logical equation it is possible to construct 64 full adder circuits. These adders are implemented with CMOS logic with technology independent optimization process and its performance are

analyzed in terms of transistor count, delay and power dissipation using Tanner EDA with TSMC MOSIS 250nm technology. From this analysis the optimized equation is selected and it is implemented in terms of multiplexers and it is incorporated in selected existing adder topologies like ripple carry adder, carry look-ahead adder, carry skip adder, carry select adder, carry increment adder and carry save adder and its performance is analyzed in terms of area (slices used) and maximum combinational path delay as a function of size.

Mathematically it is also possible to calculate the delay of a circuit by constructing delay models instead of simulation tools using logical effort methods. The logical effort provides a simple method "on the back of an envelope" [6, 7] to choose the best topology (logical constructs) and number of stages of logic for a function. An example to calculate the delay of a full adder is shown in Figure (1) using the expression $SUM = A \oplus B \oplus C$ and $CARRY = \overline{\overline{AB} \bullet \overline{AC} \bullet \overline{BC}}$. The circuit is realized as two stage network, stage1 and stage2 respectively. Assume that the input capacitance of 10pf on each input and it will drive the output capacitance with a maximum of 10pf.

Delay calculation for SUM	Delay calculation for CARRY
The Electrical effort H = 10pf/10pf (Output capacitance / Input capacitance) = 1	The Electrical effort H = 10pf/10pf (Output capacitance / Input capacitance) = 1
Path logical effort G along path (N-W) = $\prod g_1 \bullet g_2$ (where g_1 and g_2 are logical effort of XOR gate) =4X4=16	Path logical effort G along path (N-W) = $\prod g_3 \bullet g_4$ (where g_3 and g_4 are logical effort of NAND gate) =4/3X4/3 = 16/9
The parasitic delay along path (N-W) P = 4+4=8 (The parasitic delay of XOR gate is 4)	The parasitic delay along path (N-W)P =2+2=4 (The parasitic delay of NAND gate is 2)
The Branching effort B =1 [because all of the fan-out's along the path are one]	The Branching effort B =1 [because all of the fan-out's along the path are one]
Number of stages N= 2 stages	Number of stages N= 2 stages
The path Efforts F=GBH =(16)X1X1=16	The path Efforts F=GBH =(16/9)X1X1=16/9
The path delay $\hat{D} = N F^{1/N} + P$ = 2X(16)$^{1/2}$+8 =16ps	The path delay $\hat{D} = N F^{1/N} + P$ = 2X(16/9)$^{1/2}$+4 =6.6ps



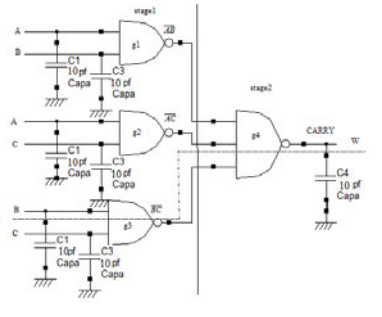


Fig. 1. Logical Delay Model for Full Adder Circuit

So the total delay will be the sum of CARRY and SUM which is equal to 22.6ps. From this observation the delay of the circuit vary with change in the input and output capacitance value.

4 Simulation and Performance Analysis of Full Adder

The proposed 20 different Boolean expressions (logic construction) are simulated using Tanner EDA with BSIM3v3 250nm technology with supply voltage ranging from 1V to 2V in steps of 0.2V. All the full adders are simulated with multiple design corners (TT, FF, FS, and SS) to verify that operation across variations in device characteristics and environment. The simulated setup for optimized full adder's (using XOR,MUX) test bed and its gate equivalent along with its input/output waveform is shown in Figure (2). The test bed is supplied with a nominal voltage of 2V in steps of 0.2V and it is invoked with the technology library file Generic 025 and it is specified with TT, FF, FS and SS conditions. The W/L ratios of both nMOS and pMOS transistors are taken as 2.5/0.25µm. To establish an unbiased testing environment, the simulations have been carried out using a comprehensive input signal pattern, which covers every possible transition for a 1- bit full adder.

The frequencies have been chosen in the range from 10 to 200MHz and its input and output capacitances are set to 10pf. The three inputs to the full adder are A, B, C and all the test vectors are generated and have been fed into the adder cell. The cell delay has been measured from the moment the inputs reach 50% of the voltage supply level to the moment the latest of the SUM and CARRY signals reach the same voltage level. All transitions from an input combination to another (total 8 patterns, 000, 001, 010, 011, 100, 101, 110, 111) have been tested, and the delay at each transition has been measured. The average has been reported as the cell delay. The power consumption is also measured for these input patterns and its average power has been reported in Table 1.

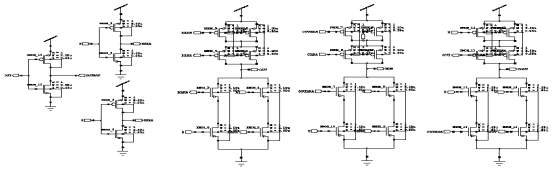


Fig. 2. A. Snap Shot of Full Adder with XOR and MUX

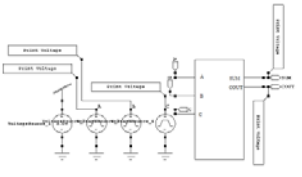


Fig. 2. B. Test bed For Full Adder

The simulation results are shown in Table 1. The performance of all the full adders has been analyzed in terms of delay, transistor count and power dissipation. It is observed that adder designed with XOR and MUX has the least delay, transistor count and power dissipation when compared to other combinations of gate. So the adder realized with MUX and XOR is considered to be the optimized adder in terms of delay, transistor count and power dissipation. The second optimized full adder is realized from XNOR, NOT and MUX.

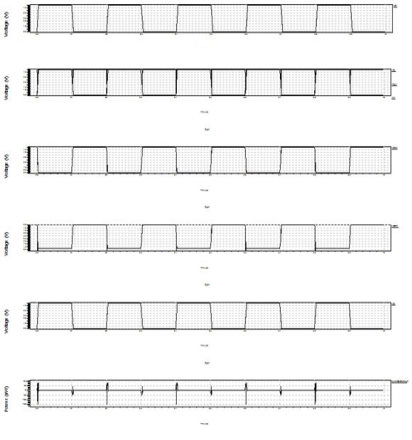


Fig. 2. C. Input/ Output Wave of Full Adder

The worst case full adder construction is not using NOR gate which occupies large transistor count, dissipates large power and has longer delay. The other optimized solution for constructing full adders are using NAND gates only, XOR, XNOR,MUX combination and XOR, AND,OR,MUX combination. Figure (3) shows the simulation result of full adders in terms of delay, transistor count (TC) and power dissipation (PWR-DISSP). The power-delay product all the full adders is shown in Figure (4).

Table 1. Simulated Result for 20 different Full adders

Full adder using	Avg Delay (ps)	Transistor count	Avg Power Dissipation(μW)
XOR,AND, OR (1)	36.1	38	12.693
XOR,AND,OR(2)	33.1	30	9.813
XNOR,AND,OR	35	38	10.371
XNOR,AND,OR,NOT	34	32	9.833
XOR, AND	36.2	30	12.35
XOR, NAND, Negative OR	30.13	30	10.547
XNOR,NAND,NOT	29.5	34	8.752
XNOR, NAND	23.01	30	9.585
XOR,NAND	23.023	30	9.485
XOR, MUX	**18.02**	**18**	**7.704**
XNOR, MUX, NOT	19.01	20	7.769
XOR, XNOR, MUX	20.8	24	8.691
XOR, AND, OR, MUX	23.6	30	9.798
XNOR,AND, OR,MUX	26.12	30	9.058
NAND	20.2	36	10.795
NOR	40.1	51	22.25
XOR,NAND,NOR,NOT	32.2	38	10.583
XNOR,NAND,NOR,NOT	31.1	38	10.1
XNOR,NOR,NOT,OR	34.1	40	11.723
XOR,NOR,NOT,OR	35	42	12.487

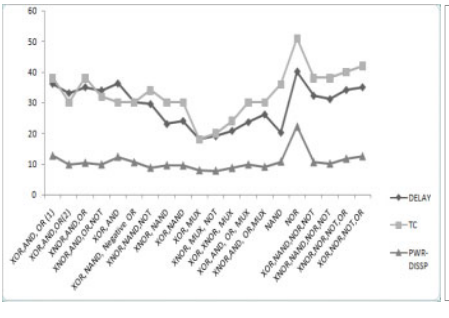


Fig. 3. Simulation result of adders in terms delay, area and power

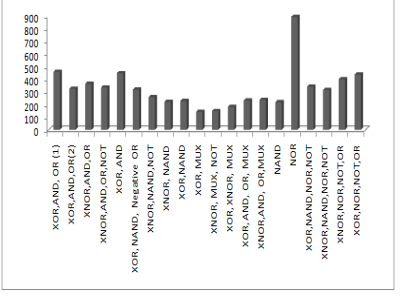


Fig. 4. Power delay product

5 FPGA Implementation

In this work the adder structures used are: Ripple Carry Adder, Carry Look-Ahead Adder, Carry Save Adder, Carry Increment adder, Carry Select Adder, Carry Skip Adder. From section IV it is observed that the optimized equation for implementing 1-bit full

adder is using XOR and MUX. So the primitive of this adder cell is implemented with multiplexer and this module is incorporated with existing adder topologies. The target FPGA device chosen for the implementation of these adders was Xilinx ISE 12.1 Spartan3E XC3S500-5FG320. Each adder type was implemented with bit sizes of: 8, 16, 32, 64 bits. This variety of sizes will provide with more insight about the performance of each adder in terms of area and delay as a function of size. Structural Gate level modeling using Verilog HDL was used to model each adder. The Xilinx ISE Foundation version 12.1i software was used for synthesis and implementation.

Table 2 contains the results obtained. The adder abbreviations used in the table are: RCA for Ripple Carry Adder, CLA for Carry Look-Ahead Adder, CSA for Carry Save adder, CIA for Carry Increment Adder, CSeA for Carry Select Adder and CSkA for Carry Skip Adder. In the Table, delay is measured in nanoseconds (ns) while area is measured in Slice Look-Up Tables (LUT) units which represent configurable logic units within the FPGA. In the Table E-A represents the adder designed with normal expression for SUM and CARRY, Op-A represents the adder topology implemented with optimized equations that are realized in terms of multiplexers. It is noticed that delay, area and power delay product are less when compared to the normal expression.

Table 2. Comparison and Results obtained for different adder topologies

Adders	**bit**	**Delay (ns)**		**Slice utilized (Area)**		**AT**		**Power Dissipation(mw)**		**PD**	
		E-A	**Op-A**	**E-A**	**Op-A**	**E-A**	**Op-A**	**E-A**	**Op-A**	**E-A**	**Op-A**
RCA	8	13.2	**12.93**	9	9	118.83	116.37	80.98	80.1	1069.18	1035.69
	16	21.69	**21.11**	18	18	390.42	379.89	81.1	79.98	1759.06	1687.98
	32	38.67	**37.46**	37	37	1430.6	1385.87	82.3	78.5	3182.13	2940.3
	64	72.62	**70.16**	74	74	5373.6	5191.77	85.67	80.2	6221.01	5626.75
CLA	8	13.2	**12.93**	9	9	118.83	116.37	78.91	77.5	1041.85	1002.08
	16	21.69	**21.11**	18	18	390.42	379.89	80.98	78.2	1756.46	1650.41
	32	38.67	**37.46**	37	37	1430.6	1385.87	81.2	79.1	3139.6	2962.77
	64	72.62	**70.16**	74	74	5373.6	5191.77	82.3	79.3	5976.3	5563.61
CSA	8	12.06	**11.1**	14	**13**	168.81	144.3	85.23	80.98	1027.7	898.878
	16	20.02	**19.8**	27	**23**	540.49	455.4	87.1	82.3	1743.57	1629.54
	32	34.94	**32.12**	55	**52**	1921.5	1670.24	88.45	82.3	3090.18	2643.48
	64	56.8	**54.23**	110	**106**	6247.9	5748.38	90.12	84.01	5118.73	4555.86
CIA	8	12.21	**11.9**	12	**11**	146.56	130.9	85.23	80.18	1040.91	954.142
	16	16.57	**14.32**	24	**22**	397.66	315.04	87.2	82.03	1444.82	1174.67
	32	27.45	**25.67**	49	**47**	1345.1	1206.49	88.23	82.03	2422.09	2105.71
	64	47.2	**45.21**	100	**98**	4719.7	4430.58	90.1	84.12	4252.45	3803.07
CSEA	8	12.6	**10.11**	15	**11**	188.94	111.188	78.91	77.5	993.95	783.37
	16	21	**15.75**	32	**30**	672.1	472.56	80.98	78.2	1700.82	1231.81
	32	37.93	**24.36**	67	**61**	2541	1486.2	81.2	79.1	3079.51	1927.19
	64	71.77	**31.41**	135	**125**	9689	8790.12	82.3	79.3	5906.67	2490.65
CSkA	8	12.78	**11.15**	13	13	166.17	144.885	78.91	77.5	1008.63	863.738
	16	23.27	**14.52**	23	**22**	535.28	319.352	80.98	78.2	1884.65	1135.15
	32	40.15	**23.26**	45	**44**	1806.9	1023.22	81.2	79.1	3260.5	1839.47
	64	49.22	**35.73**	79	**78**	3888.5	2787.1	82.3	79.3	4050.97	2833.55

6 Summary

A new low-power, high-speed full adder cell is proposed using XOR and MUX gates. Its performances have been analyzed and reported in section IV. This optimized adder is designed with fully MUX based structure in FPGA using VERILOG HDL and this module is incorporated in the existing adder topologies and its comparison is made. The comparison of delay, slice occupied, AT and its power dissipation is depicted in the Figure (5). From this analysis it is found that for all the adder topologies the delay is less when compare to the existing adder with normal equation ($SUM = A \oplus B \oplus C$ and $CARRY = AB + AC + BC$) and it is also observed that the delay for RCA and CLA are the same and its distribution is shown in the graph (Figure 5a). In case of slice utilized there is no change occurs for RCA and CLA hence its distribution is shown as single red line in the chart (Figure 5b). From AT chart (Figure 5c) it is noticed that the AT value is large for 64 bit carry select adders and adders like ripple carry adder, carry look ahead adder and carry increment adder have less AT Value. From PD distribution (Figure 5d) less power dissipation occurs for carry increment and ripple carry adders, maximum dissipation occurs for carry save and carry skip adders. According to the presented results, the adder topology which has the best compromise between area, delay and power dissipation are carry look-ahead and carry increment adders and they are suitable for high performance and low-power circuits. The fastest adders are carry select and carry save adders with the penalty of area. The simplest adder topologies that are suitable for low power applications are ripple carry adder, carry skip and carry bypass adder with least gate count and maximum delay.

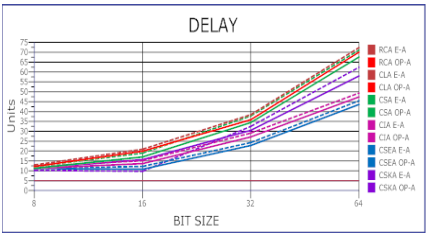


Fig. 5. A. Delay Chart

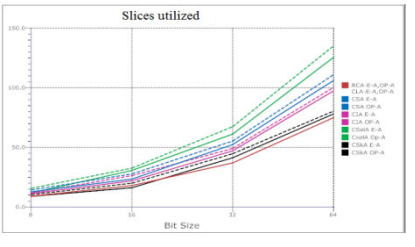


Fig. 5. B. Slices utilized Chart

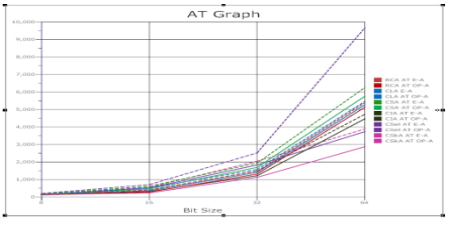


Fig. 5. C. AT Value Chart

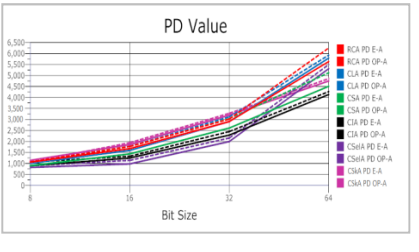


Fig. 5. D. PT Value Chart

7 Conclusion

An extensive performance analysis of 1-bit full-adder cells has been presented. Technology independent logic optimization is used to design 1-bit full adder with 20 different Boolean expressions and its performance was analyzed in terms of transistor count, delay and power dissipation using Tanner EDA with TSMC MOSIS 250nm technology. From this analysis XOR and MUX based expression provides low transistor count, minimum delay and minimum power dissipation when compared to other logic equations. The second optimized full adder can be realized using XNOR, NOT and MUX. The other optimized solution for constructing full adders are using NAND gates only, XOR, XNOR,MUX combination and XOR, AND,OR,MUX combination. The worst case full adder construction is not using NOR gate which occupies large transistor count, dissipates large power and has longer delay. Logical effort delay model to estimate the parasitic delay is also presented. Using the optimized expression the primitive adder cell is implemented with multiplexer and this module is incorporated with existing adder topologies like ripple carry adder, carry look-ahead adder, carry skip adder, carry select adder, carry increment adder and carry save adder and its performance is analyzed in terms of area (slices used) and maximum combinational path delay as a function of size. The target FPGA device chosen for the implementation of these adders was Xilinx ISE 12.1 Spartan3E XC3S500-5FG320. The comparison and its simulation results have been presented. Based on the comparison it is observed that number of slices occupied, power dissipation and delay are less using the optimized expression. The work presented in this paper gives more insight and deeper understanding of constituting modules of the adder cell to help the designers in making their choices.

References

[1] Chang, S.-C., van Ginneken, L.P.P.P.: Circuit Optimization by Rewiring. IEEE Transaction on Computers 48(9) (September 1999)
[2] Kwon, O.-H.: A Boolean Extraction Technique For Multiple-Level Logic Optimization. IEEE (2003)
[3] Uma, R.: 4-Bit Fast Adder Design: Topology and Layout with Self-Resetting Logic for Low Power VLSI Circuits. International Journal of Advanced Engineering Sciences and Technology 7(2), 197–205 (2011)
[4] Dhavachelvan, P., Uma, G.V., Venkatachalapathy, V.S.K.: A New Approach in Development of Distributed Framework for Automated Software Testing Using Agents. International Journal on Knowledge –Based Systems 19(4), 235–247 (2006)
[5] Karandikar, S.K., Sapatnekar, S.S.: Fast Comparisons of Circuit Implementations. IEEE Transaction on Very Large Scale Integration (VLSI) Systems 13(12) (December 2005)
[6] Sutherland, I., Sproull, B., Harris, D.: Logical Effort: Designing Fast CMOS Circuits. Morgan Kaufmann Publisher (1999)
[7] Uma, R., Vijayan, V., Mohanapriya, M., Paul, S.: Area, Delay and Power Comparison of Adder Topologies. International Journal of VLSI and Communication Systems (2012)

[8] Victer Paul, P., Vengattaraman, T., Dhavachelvan, P.: Improving efficiency of Peer Network Applications by formulating Distributed Spanning Tree. In: Proceedings 3rd International Conference on Emerging Trends in Engineering and Technology, ICETET 2010, Art. no. 5698439, pp. 813–818 (2010)
[9] Aguirre-Hernandez, M., Linares-Aranda, M.: CMOS Full-Adders for Energy-Efficient Arithmetic Applications. IEEE Transactions on Very Large Scale Integration (VLSI) Systems 19(4) (April 2011)
[10] Pudi, V., Sridhara, K.: Low Complexity Design of Ripple Carry and Brent Kung Adders in CA. IEEE Transactions on Nanotechnology 11(1), 105–119 (2012)
[11] Shubin, V.V.: Analysis and Comparison of Ripple Carry Full Adders by Speed. In: International Conference and Seminar on Micro/Nano Technologies and Electron Devices, EDM 2010, pp. 132–135 (2010)
[12] Veeramachaneni, S., Srinivas, M.B.: New Improved 1-Bit Full Adder Cells. IEEE (2008)

Robust Iris Templates for Efficient Person Identification

Abhishek Gangwar, Akanksha Joshi, Renu Sharma, and Zia Saquib

Center for Development of Advanced Computing, Mumbai, India
{abhishekg,akanksha,renu,saquib}@cdac.in

Abstract. Iris recognition is seen as a highly reliable biometric technology. The performance of iris recognition is severely impacted when encountering irises captured in realistic conditions. The selection of the features subset and the classification is an important issue for iris biometrics. In this paper we propose new methods for feature extraction and template creation during enrollment to improve the performance of iris recognition systems. The experiments are based on storing i) multiple templates (template group) for a user ii) Single template by taking average mean of multiple templates iii) Single template calculated from multiple templates using Direct Linear Discriminant Analysis (DLDA). We used CASIA Iris Interval database for our experiments. Experiments report significant improvement in the performance of iris recognition.

Keywords: Feature Extraction, Biometric Identification, Wavelet Transform, template creation.

1 Introduction

With increase in terrorism and illegal acts, there is a growing demand for more secure and reliable identification in our society that can replace the traditional means of identification. In order to make a decision of which biometric product or combination of products would satisfy stated requirements; different factors need to be evaluated. Factors for consideration include accuracy of a specific technology, user acceptance, and the costs of implementation and operation.

The iris is seen as a highly reliable and accurate biometric technology because each human being is characterized by unique irises that remain relatively stable over the life period. Iris is unique feature present in the form of ring around pupil of a human eye in all the human beings. Its complex pattern contains many distinctive feature such as arching ligaments, crypts, radial furrows, pigment frill, pupillary area, ciliary area, rings, corona, freckles and zigzag collarette [1, 2] which gives a unique set of feature for each human being, even irises of identical twins are different. Furthermore, the iris is more easily imaged than retina; it is extremely difficult to surgically tamper iris texture information and it is possible to detect artificial irises. The surface of the iris is composed of two regions, the central pupillary zone and the outer ciliary zone.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 255–263.
springerlink.com

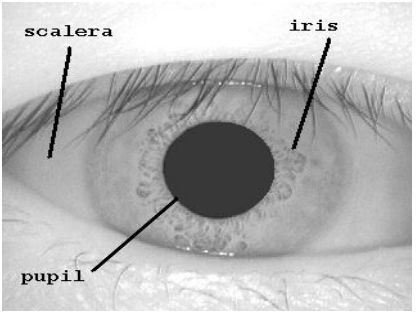


Fig. 1. Iris image from Casia database

In this paper we propose new methods for feature extraction and template creation to improve the performance of iris recognition systems in more realistic situations. The experiments are based on storing i) multiple templates (template group) for a user ii) Single template by taking average mean of multiple templates iii) Single template calculated from multiple templates using Direct Linear Discriminant Analysis (DLDA). We evaluated our approach on images from CASIA database [12]. The images are captured in near infrared illumination. First we applied wavelets on the iris region and achieved an accuracy of 74.11%, whereas we achieved significant improvement in the performance of iris recognition using templates creation methods mentioned above.

The remainder of this paper is as follows. Section 2 provides a brief overview of iris recognition, then edge detection, Hough transform, Wavelet Transform, and DLDA. Section 3 explains our proposed approach. Section 4 gives experimental evaluation & results and section 5 provides conclusion of our work.

2 Iris Recognition

Iris recognition starts with capturing the eye and localizing the iris region. Classical iris recognition systems, e.g., Daugman's and Wildes', need the users to adjust their eye positions in order to capture their irises [3]. Furthermore, existing systems require users to be close to the capturing apparatus [4]. A common observation about eye images is that the iris region is brighter than the pupil and darker than the sclera. As a result, almost all approaches to iris localization are based on the intensity gradient or edge information. These methods depend heavily on the strong intensity contrast between the pupil and iris and between the iris and sclera. Wildes [10] proposed iris segmentation by edge detection by canny filter followed by circular Hough transform. It is based on the assumption that pupillary and limbic boundaries are circular and eyelids are parabolic in shape. Since variations in the eye, like optical size of the iris, position of pupil in the iris, and the iris orientation changes from person to person, it is required to normalize the iris image, so that the representation is common to all, with similar dimensions. Normalization process involves unwrapping the iris and converting it into its polar equivalent. It is done using Daugman's Rubber sheet model [2].

The iris features of a human eye may be extracted using wavelet transform. We can use wavelets for multiresolution decomposition analysis for iris of a human eye. Wavelets are a powerful tool and have been applied earlier for image compression and texture classification. The wavelet analysis can be done by successive low pass and high pass filtering of an image. Once features are generated, comparison of the bit patterns is done to check if the two images belong to the same person. Euclidean Distance is used for this comparison.

Here is the brief overview of the techniques used in our approach.

2.1 Canny Edge Detector

Detection of iris edges includes inner (with pupil) and outer (with sclera) edges which are located by finding the edge image using the Canny edge detector [6].

The Canny detector creates binary edge map correspondent to the identified edges in the grayscale image. It starts with finding a gradient map for each image pixel. Then non-maximum suppression is applied in following manner. For a pixel im-grad(x,y), in the gradient image, and given the orientation theta(x,y), the edge intersects two of its 8 connected neighbors. The point at (x,y) is a maximum if its value is not smaller than the values at the two intersection points. Further, the hysterisis thresholding process uses two predefined values to classify some pixels as edge or non-edge. In next step, edges are recursively extended to those pixels that are neighbors of other edges and with gradient amplitude higher than a lower threshold. Now Hough transform [5] is applied to detect circles in the edge image.

2.2 Hough Transform

For every edge pixel, the points on the circles surrounding it at different radii are taken, and their weights are increased if they are edge points too, and these weights are added to the accumulator array. Thus, after all radii and edge pixels have been searched, the maximum from the accumulator array is used to find the center of the circle and its radius. The Hough transform is performed for the iris outer boundary using the whole image, and then performed for the pupil only, instead of the whole eye, because the pupil is always inside the iris.

$$H(x_c, y_c, r) = \sum_{j=1}^{n} h(x_j, y_j, x_c, y_c, r), \tag{1}$$

Where

$$h(x_j, y_j, x_c, y_c, r) = \begin{cases} 1 & \text{if } g(x_j, y_j, x_c, y_c, r) = 0; \\ 0 & \text{otherwise.} \end{cases}$$

With

$$g(x_j, y_j, x_c, y_c, r) = (x_j - x_c)^2 + (y_j - y_c)^2 - r^2. \tag{2}$$

For each edge point (x_i, y_i), this function returns zero for every parameter triple (x_c, y_c, r) that represents a circle through that point. Parameter triple (x_c, y_c, r) represents the contour of interest for which H is maximum.

2.3 Rubber Sheet Model by Daugman

The rubber sheet model [2] assigns to each point on the iris, regardless of its size and pupillary dilation, a pair of real coordinates (r, θ), where r is on the unit interval [0, 1] and θ is an angle in [0, 360°]. The remapping of the iris image I(x, y) from raw cartesian coordinates (x, y) to the dimensionless non concentric polar coordinate system (r, θ) can be represented as:

$$I(x(r, \theta), y(r, \theta)) \rightarrow I(r, \theta) \tag{3}$$

Where $x(r, \theta)$ and $y(r, \theta)$ are defined as linear combinations of both the set of papillary boundary points $(x_p(\theta)), y_p(\theta))$ and the set of collarette boundary points $(x_c(\theta)), y_c(\theta))$ as:

$$\begin{cases} x(r,\theta) = (1 - r) * x_p(\theta) + r * x_c(\theta) \\ y(r,\theta) = (1 - r) * y_p(\theta) + r * y_c(\theta) \end{cases} \tag{4}$$

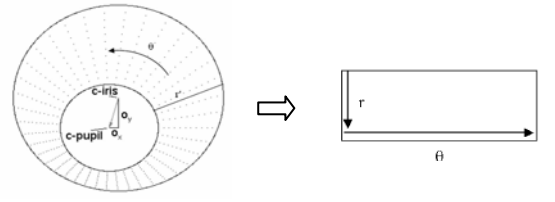


Fig. 2. Normalization process

2.4 Wavelet

Wavelets [11] are used for multiresolution analysis of signals and have been a useful tool for analyzing patterns in images. Wavelets can be applied at different scales or levels. Level one wavelet decomposition will yield four sub-images or bands; approximate, vertical, horizontal and detail. The approximate sub-band extracts the low level details from the image (LL), whereas vertical sub-band extracts the high frequencies in vertical direction and low frequencies in the horizontal direction which is basically the HL component of the sub image. The horizontal component (LH) extracts the high frequencies in horizontal direction and low frequencies in vertical direction and the diagonal part (HH) contains the high frequencies in horizontal as well as vertical direction. The second level wavelet decomposition is obtained by applying the wavelet decomposition to the approximate sub-image obtained from level one decomposition. Thus wavelet decomposition can be obtained at various levels by recursively applying the decomposition on the approximate part of the image obtained in the previous level. Two level wavelet decomposition is shown in Fig. 3.

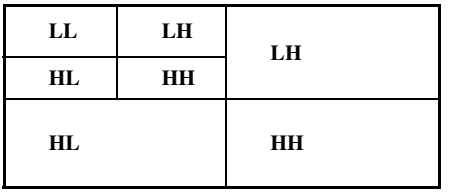


Fig. 3. Two level wavelet decomposition

2.5 Direct Linear Discriminant Analysis (DLDA)

Direct Linear Discriminant Analysis [7] is a well known classification technique and has been applied over face to reduce the feature dimensions. It basically transforms the data to lower dimensions, without losing the discrimination information. Previously, people applied PCA + LDA approach [8] but applying PCA tend to lose the discriminatory information among classes and thus yielding low accuracy. LDA aims at maximizing the ratio of between class scatter S_b and within class scatter S_w. It discards the null space of S_w, which contains the most discriminatory information according to Chen et al. [9]. The key idea of DLDA algorithm is to discard the null space of S_b, which contains no useful information rather than discarding the null space of S_w, which contains the most discriminative information. This can be done by first diagonalizing S_b and then diagonalizing S_w. The whole DLDA algorithm is outlined below.

1. First diagonalize the S_b matrix, such that

$$V^T S_b V = D \tag{5}$$

It involves finding the eigenvectors of matrix S_b and matrix D contains the corresponding eigen values. We discard those values from V which contains eigen values corresponding to 0 such that

$$Y^T S_b Y = D_b > 0 \tag{6}$$

Where Y is n*m matrix (n is the feature dimension) and contains first m columns from V and D_b is m*m matrix corresponding to non-zero eigen values.

2. Let $Z = YD_b^{-1/2}$, where

$$Z^T S_b Z = I \tag{7}$$

Where, Z unitizes S_b and reduces dimensionality from n to m.

3. Now we need to diagonalize $Z^T S_w Z$ as,

$$U^T \left(Z^T S_w Z \right) U = D_w \tag{8}$$

4. Let $A = U^T Z^T$ diagonalizes both numerator and denominator in Fisher's criteria as,

$$A^T S_w A = D_w, \; A^T S_b A = I \tag{9}$$

5. Thus, we get the final transformations as

$$T = D_b^{-1/2} AX \tag{10}$$

Where X is the feature vector extracted from the image and T is the reduced feature vector.

2.6 Euclidean Distance

The Euclidean Distance is a method to find similarity between two feature vectors. The Euclidean Distance is calculated by measuring the norm between two vectors as,

$$D = \sqrt{(X2 - X1)^2 + (Y2 - Y1)^2} \tag{11}$$

3 Proposed Approach

In our approach, we segmented the pupil and iris region using the canny edge detector and Hough transform. The extracted iris region is normalized using rubber sheet model and wavelet decomposition is applied to extract features from the iris region. We applied five level wavelet decomposition and utilized the 2nd (second) level approximate and 5th (fifth) level vertical coefficients only from the image. We have chosen the 2nd level approximate coefficients because they contain most of the information from the image. Fifth level coefficients provides the discriminatory information therefore, we also selected 5th level vertical coefficients from the image. A five level wavelet decomposition of the iris is shown in Fig. 5.

4 Experimental Results

We evaluated our approach on CASIA Iris Interval database. It contains left and right eye iris images from 249 users. We tried different combinations of wavelet (Db4) coefficients from each level and selected the second level approximate and fifth level vertical coefficients.

Experiment 1

In first experiment we took only one image to create enrollment template. We applied db4 wavelet on the iris region to extract the features. We used euclidean distance for matching of two images. We achieved an accuracy of 74.11% when we used db4 wavelet decomposition on iris region.

Experiment 2

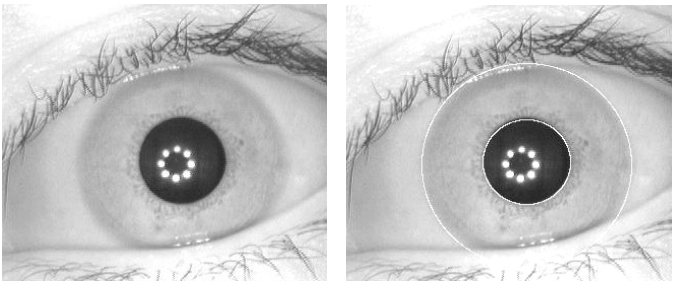
In this experiment we took multiple images (up to 5) of each user for enrollment purpose. We extracted iris features of each image using db4 wavelet. We then stored all the templates (template group) as enrollment templates for the user. We used euclidean distance for matching features. The experimental results are shown in table 1.

Experiment 3

In this experiment we took multiple images (up to 5) of each user for enrollment purpose. We extracted iris features of each image using db4 wavelet. We then calculated average mean of all the templates and created a single template to be stored as enrollment template. We used euclidean distance for matching features. The experimental results are shown in table 1.

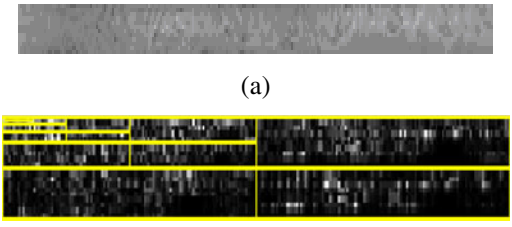
Experiment 4

In this experiment we took multiple images (up to 5) of each user for enrollment purpose. We extracted iris features of each image using db4 wavelet and then applied DLDA to create a single template to be stored as enrollment template. We used euclidean distance for matching features. The experimental results are shown in table 1.



(a) (b)

Fig. 4. (a) Image from CASIA database (b) Iris segmented image



(a)

(b)

Fig. 5. (a) Normalized Iris image (b) Five level wavelet decomposition for iris

Table 1. Identification Accuracies

Experiments	No of Images used for enrollment(gallery)				
	1	2	3	4	5
Experiment 1 (Single template)	74.11 %	-	-	-	-
Experiment 2 (Multiple templates)	74.11 %	90.00 %	97.05 %	97.64 %	98.23 %
Experiment 3 (Average mean)	74.11 %	90.00 %	90.00 %	90.58 %	91.76 %
Experiment 4 (Wavelet + DLDA)	33.52 %	38.23 %	93.52 %	94.70 %	98.82 %

We can observe from the results (given in Table 1) that there is a significant improvement in the performance of iris recognition when we increase the number of images during enrollment. In case of experiment 2, we are storing multiple templates which will increase the size of the data stored as well as the searching time. In experiment 3 and 4 we are calculating and storing only one enrollment template. It can also be observed that DLDA based approach performs well only when we have sufficient training images for the user during the enrollment. It shows that different methods require different number of training images to provide best results.

5 Conclusion

The performance of iris recognition is severely impacted in realistic conditions because of captured images encountering different illumination, pose, blur and environmental conditions etc. Experimentally we have shown that template calculated with multiple enrollment images improves the identification performance. In future we will study more robust methods to extract features from more realistic iris images and to store them in more compact form.

References

[1] Daugman, J.G.: The importance of being random: Statistical principles of iris recognition. Pattern Recognition 36(2), 279–291 (2003)
[2] Daugman, J.G.: How iris recognition works. IEEE Trans. on Circuits and Systems for Video Technology 14(1), 21–30 (2004)
[3] Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Surveys 35(4), 399–458 (2003)
[4] Daugman, J.: Biometric personal identification system based on iris analysis. U.S. Patent No. 5, 291, 560 (March 1994)
[5] Chuan Chen, T., Liang Chung, K.: An Efficient Randomized Algorithm for Detecting Circles. Computer Vision and Image Understanding 83, 172–191 (2001)
[6] Canny, J.: A Computational Approach to Edge Detection. IEEE Transaction on Pattern Analysis and Machine Intelligence 8, 679–714 (1986)
[7] Yu, H., Yang, J.: A Direct LDA Algorithm for High-Dimensional Data with Application to Face Recognition Interactive System Labs. Carnegie Mellon University, Pittsburgh
[8] Swets, D., Weng, J.: Using discriminant eigenfeatures for image retrieval. PAMI 18(8), 831–836 (1996)
[9] Chen, L., Liao, H., Ko, M., Lin, J., Yu, G.: A new lda-based face recognition system which can solve the small sample size problem. Pattern Recognition 33(10), 1713–1726 (2000)
[10] Wildes, R.P.: Iris recognition: an emerging biometric technology. Proceedings of the IEEE 85(9), 1348–1363 (1997)
[11] Frazier, M.W.: An Introduction to Wavelets through Linear algebra. Springer (1999)
[12] CASIA Iris Image Database, http://www.sinobiometrics.com

Thesaurus Based Web Searching

K.V.N. Sunitha and A. Sharada

CSE Dept, G. Narayanamma Institute of Technology and Science,
Shaikpet, Hyderabad, India
k.v.n.sunitha@gmail.com,
sharada.nirmal@gmail.com

Abstract. Search engine technology has become quite popular to help users seek information available on the web. The success of a searching system is determined by the quality and efficiency of the search results. There may be very good items on the search topic in other languages, but, search engine will generally retrieve items of only one language. Most of these search engines use pattern search which is not efficient. In this paper we present a tool that addresses this problem. Here we discuss the work carried out in developing an efficient tool that retrieves all the items of the database relevant to search term, not just the term matching. This tool retrieves all the synonym matches from both Telugu and English languages.

Keywords: Telugu dictionary, Foreign key, EngTelMap, context resolution, reverse mapping.

1 Introduction

The growth of the Web leads to high popularity of the online search services. The success of a searching system is determined by the quality and efficiency of the search results. Most users have been trained to become accustomed to the traditional search interface where a user submits a query to an input textbox. It is not easy for Web users to specify their search intentions by term combination. Sometimes a user can not generate a query correctly even though the user is clear about what to search.. The vastness of knowledge available on the WWW is the cause of our ever-increasing vulnerability to "not the best" knowledge available. As the number of indexable pages on the Web exceeds, it becomes more and more difficult for search engines to keep an up-to-date and comprehensive search index, resulting in low precision and low recall rates. Users often find it difficult to search for useful and high-quality information on the Web using general-purpose search engines, especially when searching for information on a specific topic or in a language other than English[1]. Many domain specific or language specific search engines have been built to facilitate more efficient searching in different areas. But for Indian languages, it is still not much advanced. For Telugu language, perhaps, there is no efficient search engine that is based on thesaurus.

Telugu has a vast and rich culture and literature dating back to many centuries[2]. Yet there is no widely available electronic thesaurus till date. However bilingual dictionaries are available. For NLP applications, Thesaurus is to be constructed from

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 265–272.
springerlink.com

these dictionaries. The existing searching system gives results only for the search term given and it doesn't provide any specific meanings for the given word depending upon the context.

In this paper, we present our work in designing and implementing a software tool that addresses this problem. We will focus on the architectural design of the tool. The rest of the paper is organized as follows. Section 2 reviews related work in search engine development. Section 3 discusses our research objective. In Section 4, we present our proposed system and Section 5 presents conclusion and sample results.

2 Literature Survey

There are many free software tools that provide all of the components of a search engine. Although these toolkits provide integrated environments for users to build their own domain-specific search engines, most of these tools only work for English documents and are not able to process non-English documents, especially for non-alphabetical languages. Only a few of them, such as GreenStone, support multilingual collection building. As a result, most of these tools cannot be used to create digital library for non-English collections. Another problem is that many of these tools do not provide enough technical details, and their components and building steps are tightly coupled. As a result, users often find it difficult to customize the tools or reuse the intermediate results in other applications (such as document classification) even if they have strong technical skills. The popular search engine, Google, uses the random walk algorithm, which ranks the documents according to the link structure, coupled with the local query specific score to give the final rank to the page[4].

Search engines like oingo.com, excite.com and simpli.com also provide meaning based searching. Launched in October 1999, Oingo has already introduced three fully functional products: DirectSearch, DomainSense and AdSense. DirectSearch, a meaning-based search technology, uses the company's ontology to provide more precise and effective search results. DomainSense, Oingo's meaning-based domain name suggestion technology, currently increases domain name sales for leading registrars around the world. AdSense serves the most highly targeted advertisements on the Internet; effectively targeting advertisements based on search meanings rather than keywords[3].

3 Proposed System

In our searching system results will be displayed for all the semantically related words to the given search term.

a. Synonym Based Searching

There is no controlled vocabulary or list of standardized terms or descriptors for the Web. Because there is no controlled vocabulary, the use of synonyms and variations of keywords to describe the search query is very important.

b. Context Based Searching

If the word has different meanings, depending upon the context all the different meanings will be displayed.

c. Cross Linear Searching

For a given term in Telugu we can retrieve words in English. Searching is possible for both the languages i.e searching is possible from English–Telugu as well as from Telugu–English.

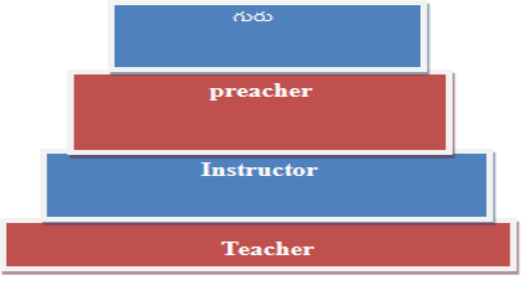


Fig. 1. Cross linear Searching for the word "Teacher"

3.1 Objectives

The objectives of this work is as under

- Manual Construction of Telugu Thesaurus
- Efficient creation and manipulation of the database for the thesaurus.
- Creating transliterable input control so that user can give input in both Telugu and English.
- Using database for context resolution
- Fetching the results using the URL from web server
- Performing web content mining for retrieving the links
- Displaying list of Websites that matches the given search criteria.

4 Architecture

4.1 Thesaurus Building

In the proposed system we use relational database(tables) to store and retrieve large amounts of data. It contains two main tables: Engwords(table for English words), Telwords(table for Telugu words).

i) EngWords Table:

This table contains all the English words. The table 'EngWords', shown in Table 1, has two attributes: 'Word' and 'Id'. 'Id' is the primary key, which uniquely identifies the words in the table. 'Word' attribute is also unique, which doesn't allow having duplicate copies of word in the table. For each word in the table a unique id is assigned so that, using the id, words can be retrieved from other table. Both will act as the candidate keys. 'Id' of TelWords acts as foreign key for EngWords table. This foreign key is used in EngTelMap as discussed in section 4.4.

Table 1. English table - EngWords

Word	Id
Temple	1
Earth	2
Smile	3
Education	4

ii) TelWords Table:

This table contains all Telugu words and its corresponding English word Id. Just like EngWords table, it also has two attributes 'Word' and 'Id' with a difference that in this table, only the attribute 'Word' is unique and hence is the primary key. Id stores the value of EngWord Id. So it may be duplicated. For example, consider word 'Earth'. It's Id in EngWords is 2. All its semantically equivalent words of Telugu will have the same value, i.e. 2, in their Id field as shown in Table 2.

Table 2. Telugu table - TelWords

Word	Id
గుడి	1
దేవాలయము	1
మందిరము	1
కోవెల	1
ఆలయము	1
భూమి	2
ధరణి	2
పృథ్వి	2
వసుధ	2

4.2 Mapping

Mapping is the main module that retrieves semantically equivalent words from Telugu if the search term is in English and vice versa. As stated above, id of EngWords table is the foreign key of TelWords table. For example, if the search term is 'earth' whose id is '2' in EngWords, the tool searches for value '2' in Id column of TelWords and retrieves all the rows of id '2' to get the respective synonyms for English word in EngWords. Thus, by this mapping we can get multiple synonyms for given English word as shown in Fig.2

Earth= భూమి(BU-mi), ధరణి(dha-ra-Ni), పృథ్వి(pru-dhvi), వసుధ(va-su-dha)

Fig. 2. Mapping word with its synonyms

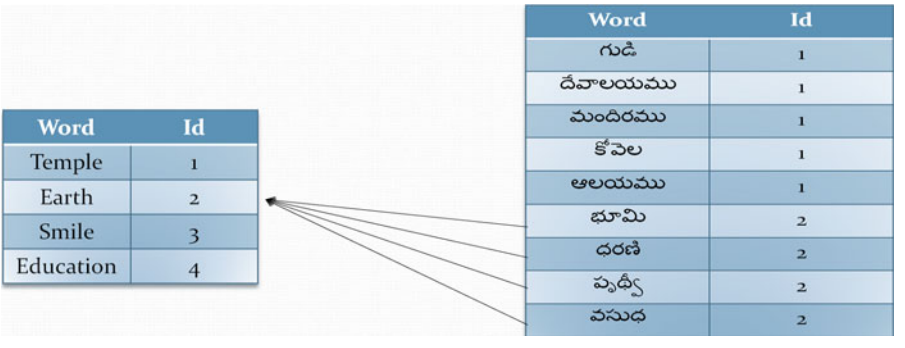


Fig. 3. Reverse mapping

Fig 3 shows reverse mapping, in which we can get English synonym for given Telugu word. So, if you give any of the above mentioned Telugu word, it will map to the English word 'Earth' with the help of Id field.

భూమి / ధరణి / పృథ్వీ / వసుధ = Earth

In fig.4, we can observe, mapping within the table, which gives multiple Telugu synonyms for given Telugu word. Given a word in Telugu, get the corresponding Id of it and then scan the whole table for that Id in Id column and extract all the words having that Id. Then the result set contains all the synonyms for the given word.

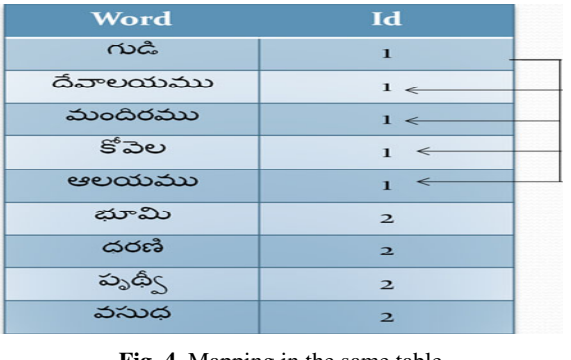


Fig. 4. Mapping in the same table

For example, consider the word 'కోవెల', Id of it is'1', searching the whole table for Id 1, we get the words like: గుడి, దేవాలయము, మందిరము, ఆలయము. Thus in this kind of mapping we can get English/Telugu synonyms for given words of same language.

5 Sample Results and Conclusion

The overall quality of web searching system is determined not only by the prowess of its searching algorithm, but also by the caliber of its corpus, both in terms of comprehensiveness (e.g. coverage of topics, language, etc.) and refinement (e.g. freshness, avoidance of redundancy, etc.).The relative corpus size estimates competitive marketing advantage and bragging rights in the context of the web searching.

We have collected 700 Telugu words with their respective synonyms in general context.

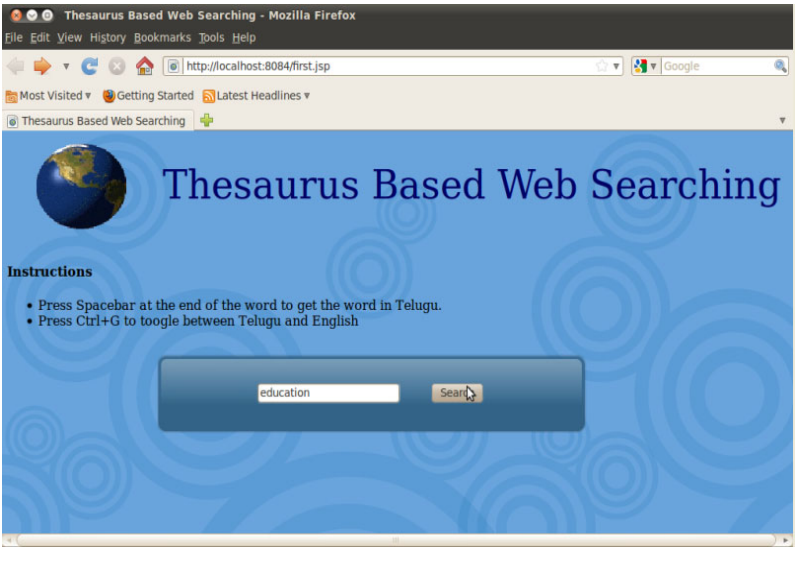


Fig. 5. Searching for the word "education"

Manual procedures of thesaurus building can be a bottleneck of our proposed approach. In our future work we are going to address the problem by making use of automated lexical acquisition. The automatically constructed thesaurus can also be taken as a starting point for developing a better searching system. In the present system we have developed tool for searching Telugu and English words. This is a starting step in construction of thesaurus for Telugu, which is a very rich language but there is a huge scarcity of resources. In our future work we would like to extend this tool to context based multi lingual web search where different meanings of the same word are also considered.

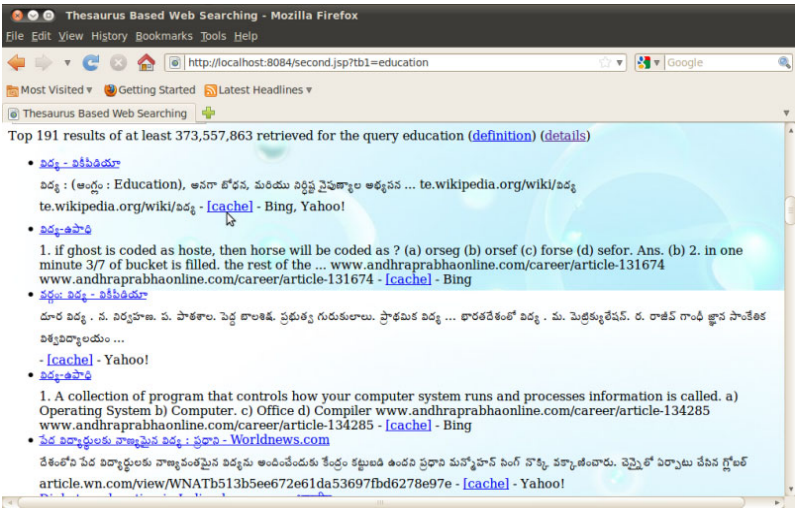


Fig. 6. Search results for the word “education” in Telugu

Fig. 7. Search results for the word “education” in Telugu & English

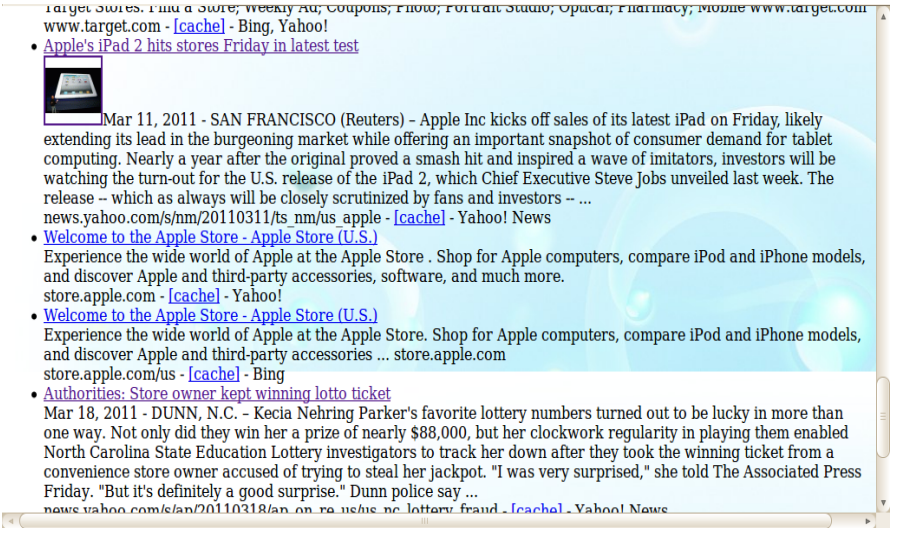


Fig. 8. Few more results for the search term

References

1. Chau, M., et al.: SpidersRUs: Creating specialized search engines in multiple languages. Science Direct, Decision Support Systems 45, 621–640 (2008)
2. Uma Maheshwara Rao, G.: Morphological complexity of Telugu. In: ICOSAL-2 (2000)
3. Bhattacharya, P., et al.: A multi Lingual Meaning Based Search Engine
4. Brin, S., Page, L.: The anatomy of a Large-Scale Hypertextual Web Search Engine. In: Proceedings of the7th WWW Conference, Brisbane, Australia (April 1998)
5. Chen, H., Chau, M., Zeng, D.: CI Spider: a tool for competitive intelligence on the web. Decision Support Systems 34(1), 1–17 (2002)
6. Lovins, J.: Development of stemming algorithm. Journal of Mechanical Translation and Computational Linguistics 11, 22–31 (1968)
7. Sunitha, K.V.N., Sharada, A.: Building an Efficient Language Model based on Morphology for Telugu ASR. In: KSE-1, CIIL, Mysore (March 2010)
8. Sunitha, K.V.N., Sharada, A.: Telugu Text Corpora Analysis for Creating Speech Database. IJEIT 1(2) (December 2009) ISSN 0975-5292
9. Paice, C., Husk, G.: Another Stemmer. ACM SIGIR Forum 24(3), 566 (1990)
10. Porter, M.F.: An algorithm for suffix stripping. In: Readings in Information Retrieval, pp. 313–316. Morgan Kaufmann Publishers Inc., San Francisco (1997)
11. Xu, J., Bruce Croft, W.: Corpus based stemming using co-occurrence of word variants. ACM Trans. Inf. Syst. 16(1), 61–81 (1998)
12. Dawson, J.L.: Suffix removal for word conflation. Bulletin of the Association for Literary and Linguistic Computing 2(3), 33–46 (1974)

An Adaptive Design Pattern for Genetic Algorithm Based Autonomic Computing System

B. Naga Srinivas Repuri[1], Vishnuvardhan Mannava[1], and T. Ramesh[2]

[1] Department of Computer Science and Engineering,
K L University, Vaddeswaram, 522502, A.P., India
bhaskarr1205@gmail.com, vishnu@kluniversity.in
[2] Department of Computer Science and Engineering,
National Institute of Technology, Warangal, 506004, A.P., India
rmesht@nitw.ac.in

Abstract. The need for adaptability in software is growing, driven in part by the emergence of pervasive and Autonomic computing. In many cases, it is desirable to enhance existing programs with adaptive behavior, enabling them to execute effectively in dynamic environments. Increasingly, software systems should self adapts to satisfy new requirements and environmental conditions that may arise after deployment. Due to their high complexity, adaptive programs are difficult to specify, design, verify, and validate. In this paper we propose an approach for Genetic Algorithm based Autonomic System using Design Patterns. The proposed pattern satisfies the properties of Autonomic System. This pattern is an amalgamation of different Design Patterns like Observer, Strategy, Singleton, Adapter and Thread per connection. For monitoring we use Observer pattern, Decision making we use Strategy Design Pattern. Adapter, Singleton and Thread per Connection pattern are used for execution of Genetic Algorithm population. Proposed pattern distribute population of Genetic Algorithm to different clients for execution. Main objective of the proposed system is to reduce the load of the server. This design pattern solve multi objective optimization problem using Genetic Algorithm. The pattern is described using a java-like notation for the classes and interfaces. A simple UML class and Sequence diagrams are depicted.

Keywords: Design Patterns, distributed system, Genetic Algorithms and Autonomic System.

1 Introduction

A Genetic Algorithm (GA) is a problem solving method inspired by Darwin's theory of evolution: a problem is solved by an evolutionary process resulting in a best (fittest) solution (survivor). In a GA application, many individuals derive, independently and concurrently, competing solutions to a problem. These solutions are then evaluated for fitness and individuals survive and reproduce based upon their fitness. Eventually, the best solutions emerge after generations of evolution.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 273–282.
springerlink.com

The flow of a typical GA simulation is as follows: First, a GA server creates many individuals randomly. Each of these individuals is tested for fitness. Based on their fitness, measured by a fitness function that quantifies the optimality of a solution, the server selects a percentage of the individuals that are allowed to crossover with each other, analogous to gene sharing through reproduction in biological organisms. The crossover between two parents produces offspring, which have a chance of being randomly mutated. A child thus produced is then placed into the population for the next generation, in which it will be evaluated for fitness. The process of selection, crossover, and mutation repeats until the new population is full and the new generation repeats the behavior of the previous generation. After many generations, the individuals are expected to become more adept at solving the problem to which the GA is being applied.

In order for a GA simulation to work well, there needs to be a significant number of individuals within a population, and the simulation needs to be allowed to run for many generations. Furthermore, the simulation will typically need to be run repeatedly while parameters such as mutation rate, population size and crossover functions are tuned. Thus, a successful GA simulation requires the calculation of the fitness function thousands of times or more. It is therefore critical that the function that performs the calculation of the fitness, called the fitness function, can be executed as speedily as possible.

Design Patterns have, over the last decade, fundamentally changed the way we think about the design of large software systems. Using Design Patterns not only helps designers exploit the community's collective wisdom and experience as captured in the patterns, it also enables others studying the system in question to gain a deeper understanding of how the system is structured, and why it behaves in particular ways. And as the system evolves over time, the patterns used in its construction provide guidance on managing the evolution so that the system remains faithful to its original design, ensuring that the original parts and the modified parts interact as expected. Although they are not components in the standard sense of the word, patterns may, as has been noted, be the real key to reuse since they allow the reuse of design, rather than mere code. But to fully realize these benefits, we must ensure that the designers have a thorough understanding of the precise requirements their system must meet in applying a given pattern, as well as automated or semi-automated ways of checking whether the requirements have been satisfied. To that end, the work we present in describes an approach to specifying Design Patterns precisely using formal contracts. Our goal in this paper is to extend that work, and to develop a runtime monitoring approach that allows system designers to determine whether the patterns used in constructing a system have been applied correctly. We use an aspect-oriented programming approach to achieve this goal.

As distributed computing applications grow in size and complexity in response to increasing computational needs, it is increasingly difficult to build a system that satisfies all requirements and design constraints that it will encounter during its lifetime. Many of these systems must operate continuously, disallowing periods of downtime while humans modify code and fine-tune the system. For instance, several studies document the severe financial penalties incurred by companies when facing problems such as data loss and data inaccessibility. As a result, it is important for applications to be able to self-reconfigure in response to changing requirements and environmental

conditions. IBM proposed Autonomic computing as a means for automating software maintenance tasks. Autonomic computing refers to any system that manages itself based on a system administrator's high level objectives while incorporating capabilities such as self-reconfiguration and self-optimization. Typically, developers encode reconfiguration strategies at design time, and the reconfiguration tasks are influenced by anticipated future execution conditions. We propose an approach for incorporating Genetic Algorithms as part of the decision-making process of an Autonomic System. This approach enables a decision making process to dynamically evolve reconfiguration plans at run time. Figure 1 demonstrates architecture of Autonomic System.

Proposed system solves multi objective optimization problem using Genetic Algorithm. It optimize the memory management problem, Genetic Algorithm will generate population based on input stream provided by the user. Server dynamically pick appropriate fitness function for Genetic Algorithm, Genetic Algorithm generate population for finding solution. Our Proposed pattern distributes population to different clients for evaluation. Main objective of proposed pattern is to reduce load of Genetic Algorithm server.

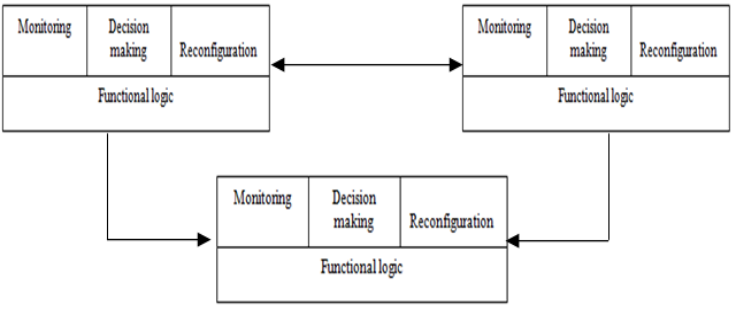


Fig. 1. Autonomic System

2 Related Work

In this section we present some works that deal with different aspects of Autonomic Systems and their design. Nick Burns and Mike Bradley paper[1] discuss applying Genetic Algorithm for distributing computing we take this paper is base paper here we are applying Genetic Algorithms and Design Patterns in Autonomic Systems. The author of the paper uses compostie, Singleton half-sys and half-asyn patterns for system designing. In Jason O. Hallstrom and Neelam Soundarajan[2] uses Observer pattern for monitoring approach for determining whether the pattern contracts used in developing a system are respected at runtime. In Andres J. Ramirez and David B. Knoester [3] proposes applying Genetic Algorithms for decision making in Autonomic computing.

In Andres J. Ramirez and David B. Knoester[4] uses Distributed Adapters Pattern (DAP) in the context of remote communication between two components for object oriented applications. In this paper uses all base paper concepts for designing new system for Autonomic computing. Genetic Algorithms also have been used to design

overlay multicast networks for data distribution [9]. These overlay networks must balance the competing goals of diffusing data across the network as efficiently as possible while minimizing expenses. A common approach for integrating various objectives in a Genetic Algorithm is to use a cost function that linearly combines several objectives as a weighted sum [12]. Although most of these approaches [10] achieved rapid convergence rates while producing overlay networks that satisfied the given constraints, to our knowledge, the methods were not applied at run time to address dynamic changes in the network's environment [11]. V.S.Prasad Vasireddy, Vishnuvardhan Mannava, and T. Ramesh paper [8] discuss applying an Autonomic Design Pattern which is an amalgamation of chai n of responsibility and visitor patterns that can be used to analyze or design self-adaptive systems.

3 Proposed Autonomic Design Pattern

In this paper we propose a pattern for autonomic computing system that solve multi objective optimization problem using Genetic Algorithms. Main objective of proposed Design Pattern is to reduce work load of Genetic Algorithm server. For reducing the server load, we proposed an approach that is distributing load (population of Genetic Algorithm). Proposed pattern is an amalgamation of different Design Patterns such as Observer, Strategy, Singleton, thread per connection and Adapter Design Pattern.

In this pattern server will take input from user and based on the input, server chooses fitness function using Strategy pattern. Fitness function is used for generating population in Genetic Algorithm, based on fitness function Genetic Algorithm will generate population. Each population in Genetic Algorithm will provide different solutions to the problem. This population is distributed to different clients for evaluation. For distribution we use Thread per connection Design Pattern. Each client in the proposed system executes Singleton Design Pattern. Singleton Design Pattern used to restrict clients from execution of single population at a time.

After the evaluation of the population, clients send results to server. Using Adapter Design Pattern serer will convert the results to specific format that is understood by the server. Out of all possible results server will choose appropriate result. Proposed pattern satisfies the properties of Autonomic System.

4 Proposed Autonomic Design Pattern

To facilitate the organization, understanding, and application of the proposed Design Patterns, this paper uses a template similar in style to that used in [10].

4.1 Pattern Name

A Novel Adaptive Design Pattern for Genetic Algorithm Based Autonomic System.

4.2 Classification

Structural-Decision-Making

4.3 Intent

Systematically applies the Design Patterns to an Autonomic System for solving Genetic Algorithm based multi objective optimization problem.

4.4 Proposed Pattern Structure

A UML class diagram for the proposed Design Pattern can be found in Figure 2.

4.5 Participants

(a) **Input stream:** input stream supplies problem to system, input stream supplies problem to server class server class will try to find the solution for the problem
(b) **Server:** Server will take input from the input stream, based on the input it will find the fitness function with the help of the Strategy pattern based on the fitness function it will find the possible chromosomes for the problem. Each chromosome will distributed to different client and take result from target Adapter manipulate according to the server pattern finally found the optimum results out of all possible result.
(c) **Observer:** it will update the classes based on the results of the different client finally return all the values to server.
(d) **Adapter:** it will convert the results of the clients to server pattern; results of the clients are different from server pattern then target Adapter will convert according to the server pattern.
(e) **Server thread:** server Thread will create new Thread for every client and assign the chromosomes to the clients.
(f) **Concrete Strategy:** concrete Strategy will consists of fitness functions based on the input stream it will choose the appropriate concrete Strategy class it will provide the fitness function to the server.
(g) **Concrete server:** stores state of interest to concrete server objects. Sends a notification to its Observers when its state changes. Concrete
(h) **Concrete Observer:** maintains a reference to a Concrete Server. Stores state that should stay consistent with the subject's. Implements the Observer updating interface to keep its state consistent with the subject's.

4.6 Related Design Patterns

The intent of the adaptive Design Pattern Design Pattern is similar to the Configuration pattern. The Configuration pattern decouples structural issues related to configuring services in distributed applications from the execution of the services themselves. The **Configuration pattern [1]** has been used in frameworks for configuring distributed systems to support the construction of a distributed system from a set of components. In a similar way, the Adaptive Design Pattern decouples service initialization from service processing. The primary difference is that the Configuration pattern focuses more on the active composition of a chain of related services, whereas the Adaptive Design Pattern focuses on the dynamic initialization of service handlers at a

particular endpoint. In addition, the Adaptive Design Pattern focuses on decoupling service behavior from the service's concurrency strategies.

The **Manager Pattern [7]** manages a collection of objects by assuming responsibility for creating and deleting these objects. In addition, it provides an interface to allow clients access to the objects it manages. The Adaptive Design Pattern can use the Manager pattern to create and delete Services as needed, as well as to maintain a repository of the Services it creates using the Manager Pattern. However, the functionality of dynamically configuring, initializing, suspending, resuming, and terminating a Service created using the Manager Pattern must be added to fully implement the Adaptive Pattern.

4.7 Roles of Our Design Patterns

(a) **Strategy:** Strategy Design Pattern will choose the fitness function based on the input stream. Fitness function will wary based on the input stream. It will choose appropriate fitness function that is suitable for Autonomic System. Strategy Design Pattern will use for decision making for Autonomic System [10].

(b) **Observer:** Observer Design Pattern will use for monitoring in Autonomic Systems. Observer will monitor the clients' responses and note the observations. Monitoring helps to reduce the server time (waiting for the client response). If a client is unable to produce result then Observer will inform the details of the client to the server [10].

(c) **Singleton:** Singletons Design Pattern will use execute the population in client. The main objective of the Singleton in this Autonomic System is each client will take only one population for execution [3].

(d) **Thread per Connection:** model, is applied to the interaction between the clients and the server. For each client, the server spawns a separate Server Thread dedicated to interacting with the client [10].

(e) **Target Adapter** will take results from client it convert result as per the server pattern if it is server pattern no need of conversion of result otherwise it will be changed [12].

4.8 Applicability

Use the Autonomic System using Design Pattern when

- The Strategy chooses the reconfiguration plan based on the input stream of the modules.
- You need different variants of an algorithm. For example, you might define algorithms reflecting different space/time trade-offs. Strategies can be used when these variants are implemented as a class hierarchy of algorithms [8].
- If the Strategy will store different fitness functions updating the fitness function are also possible in Strategy. It reduces the workload of server; the system is suitable for distributed computing [5].

5 Interfaces Definition for the Pattern Entities

Input stream:

```
Public class Inputstream
{
        public String Read()
        {
                return "0";
        }
}
```

Server:

```
Public class Server
{
      Public fitnessfunction()
      {
       i.concereteImpl();
      }
      Public int enqueuejob(){…}
      Public int jobqueue(int){..}
      Public int result(){..}
      Public getjob(){….}
}
```

Client:

```
Public class Client
{
        Public int dojob() {….. }
}
```

Observer:

```
Public class Viewone
{
      Public int update(object){..}
}
```

Strategy:

```
Public class ConcreteStrategyAlpha implements Strategy
{
Public Elaborate(String) {....}
}
Public class ConcreteStrategy implements Strategy
{
  Public void Elabrate(String) {...}
 }
```

Adapter:

```
Public class Adapter
{
   Public int resultAdapter(int ){...}
}
```

Server thread:

```
Public class serverthread
{
      Public void server(int s)
      {
       Threadt=new thread();
       t.client(int                    s);
      }
}
```

Singleton:

```
Class Singleton
 {
        private static Singleton instance = null;
      public static instance()
      {
        if( instance == null )
        {
         instance = new Singleton();
        }
        return instance;
      }
```

The view of our proposed Design Pattern can be seen in the form of a Class Diagram see Figure 2.

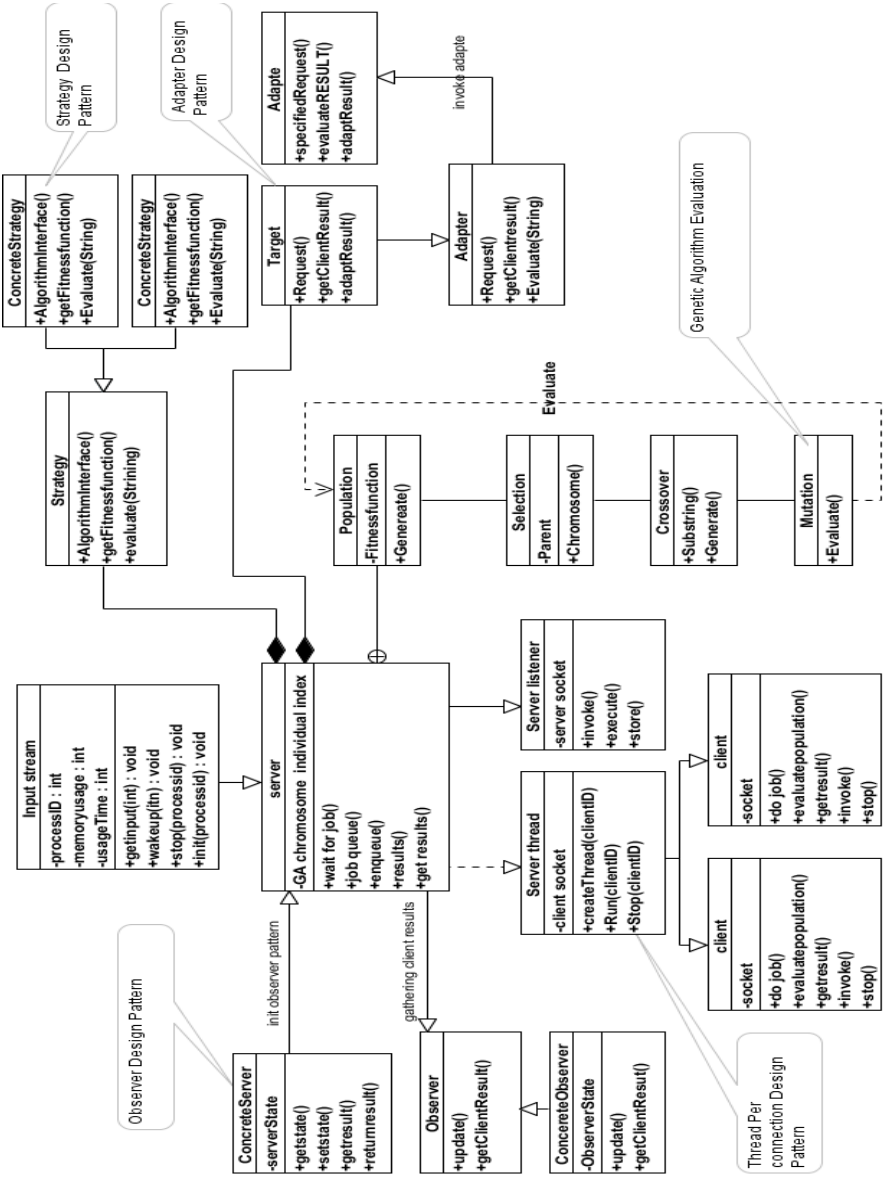


Fig. 2. Class Diagram for proposed Design Pattern

6 Case Study

To demonstrate the efficiency of the pattern we took the profiling values using the Net beans IDE and plotted a graph that shows the profiling statistics when the pattern is applied and when pattern is not applied. This is shown in figure 3. Here X-axis represents the runs and Y-axis represents the time intervals in milliseconds. Below simulation shows the graphs based on the performance of the system if the pattern is applied then the system performance is high as compared to the pattern is not applied.

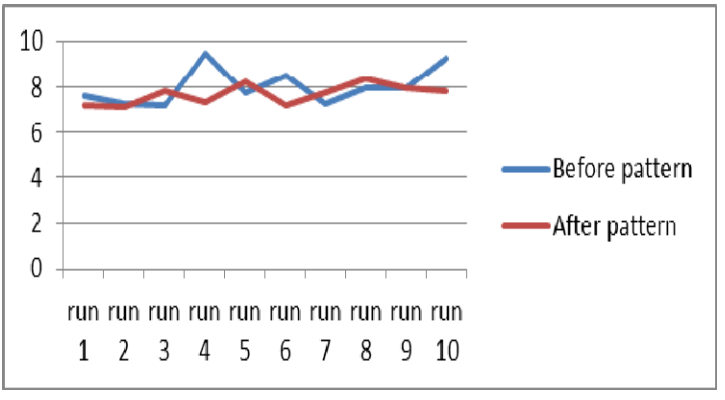


Fig. 3. Profiling statistics before applying Pattern and after applying Pattern

7 Conclusion and Future Work

In this paper we propose an approach for Autonomic computing system using Design Pattern. The system will satisfy all properties of the Autonomic System except reconfiguration. Main objective of proposed system is to reduce work load of Genetic Algorithm server by distributing Genetic Algorithm population to different clients. Each client evaluates population of Genetic Algorithm and sends back the results to Genetic Algorithm server. Server will choose best result out of all possible results from clients. Our pattern is an amalgamation of five Design Patterns those are Strategy, Observer, Singleton, Thread per connection and Adapter Design Patterns. Proposed system Observer pattern used for monitoring and Strategy pattern used for decision making.

Future work: We are trying to develop a system that will distribute the Genetic Algorithm to different clients, and to design Autonomic System. It will reconfigure based on Autonomic changes in the system using Design Patterns.

References

1. Bradley, N.B.M., Liu, M.-L.L.: Applying Design Patterns in Distributing a Genetic Algorithm Application. In: Proceedings of the International Conference on Software Engineering Research and Practice, SERP, Las Vegas, Nevada, USA, vol. 1 (2005), doi:10.1.1.86.1708
2. Ramirez, A.J., Knoester, D.B., Cheng, B.H.C., McKinley, P.K.: Applying Genetic Algorithms to Decision Making in Autonomic Computing Systems. In: 6th International Conference in Autonomic Computing, ICAP. ACM (2009), doi:10.1145/1555228.1555258
3. Hallstrom, J.O., Soundarajan, N., Tyler, B.: Monitoring Design Pattern Contracts. In: 8th International Conference on Software Engineering Knowledge Engineering, SEKE. ACM, San Francisco (2006)
4. Alves, V., Borba, P.: Distributed Adapters Pattern: A Design Pattern for Object-Oriented Distributed. ACM (2005), doi: 10.1.1.20.45 -19
5. Beck, K., Coplien, J., Crocker, R., Dominick, L., et al.: Industrial Experience with Design Patterns. In: ICSE. ACM (1996), doi:10.1.1.30.8708
6. Mannava, V., Ramesh, T.: A Novel Event Based Autonomic Design Pattern for Management of Webservices. In: Wyld, D.C., Wozniak, M., Chaki, N., Meghanathan, N., Nagamalai, D. (eds.) ACITY 2011. CCIS, vol. 198, pp. 142–151. Springer, Heidelberg (2011)
7. Mannava, V., Ramesh, T.: A Novel Adaptive Monitoring Compliance Design Pattern for Autonomic Computing Systems. In: Abraham, A., Lloret Mauri, J., Buford, J.F., Suzuki, J., Thampi, S.M. (eds.) ACC 2011, Part I. CCIS, vol. 190, pp. 250–259. Springer, Heidelberg (2011)
8. Ramirez, A.J.: Design Patterns for Developing Dynamically Adaptive Systems. Master's thesis, Michigan State University, East Lansing, Michigan (2008), doi:10.1145/1808984.1808990
9. Pree, W.: Design Patterns for Object-Oriented Software Development. Addison-Wesley, Reading (1994)
10. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns. Addison-Wesley, Reading (1994)
11. Crane, S., Magee, J., Pryce, N.: Design Patterns for Binding in Distributed Systems. In: The OOPSLA 1995 Workshop on Design Patterns for Concurrent, Parallel, and Distributed Object-Oriented Systems, Austin, TX. ACM (1995), doi:10.1.1.39.7601
12. Cheng, S.W., Garlan, D., Schmer, B.: Architecture-based self adaptation in the presence of multiple objectives. In: International Workshop on Self-Adaptation and Self-Managing Systems. ACM, New York (2006), doi:10.1145/1137677.1137679

Cross-Layer Design in Wireless Sensor Networks

S. Jagadeesan and V. Parthasarathy

Department of Computer Science and Engineering, Chettinad College of Engineering & Technology, Karur, Tamilnadu, India
jagavasan@gmail.com, sarathy.vp@gmail.com

Abstract. The literature on cross-layer protocols, protocol improvements, and design methodologies for wireless sensor networks (WSNs) is reviewed and taxonomy is proposed. The communication protocols devised for WSNs that focus on cross-layer design techniques are reviewed and classified based on the network layers they aim at replacing in the classical open system interconnection (OSI) network stack. Furthermore, systematic methodologies for the design of cross-layer solution for sensor networks as resource allocation problems in the framework of non-linear optimization are discussed. Open research issues in the development of cross-layer design methodologies for sensor networks are discussed and possible research directions are indicated. Finally, possible shortcomings of cross-layer design techniques such as lack of modularity, decreased robustness, difficulty in system enhancements, and risk of instability are discussed, and precautionary guidelines are presented.

Keywords: Cross Layer Design, Routing, IDS, Sensor Network.

1 Introduction

There exist exhaustive amount of research to enable efficient communication in wireless sensor networks (WSNs) (Akyildiz 2002). Most of the proposed communication protocols improve the energy efficiency to a certain extent by exploiting the collaborative nature of WSNs and its correlation characteristics. However, the main commonality of these protocols is that they follow the traditional layered protocol architecture. While these protocols may achieve very high performance in terms of the metrics related to each of these individual layers, they are not jointly optimized to maximize the overall network performance while minimizing the energy expenditure. Considering the scarce energy and processing resources of WSNs, joint optimization and design of networking layers, i.e., cross-layer design stands as the most promising alternative to inefficient traditional layered protocol architectures.

Accordingly, an increasing number of recent works have focused on cross-layer development of wireless sensor network protocols. In fact, recent papers on WSNs (Fang 2004)(van Hoesel 2004) (Vuran 2005) reveal that cross-layer integration and design techniques result in significant improvement in terms of energy conservation. Generally, there are three main reasons behind this improvement. First, the stringent energy, storage, and processing capabilities of wireless sensor nodes necessitate such an approach.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 283–295.
springerlink.com

The significant overhead of layered protocols results in high inefficiency. Moreover, recent empirical studies necessitate that the properties of low power radio transceivers and the wireless channel conditions be considered in protocol design (Ganesan 2002)(Zuniga 2004). Finally, the event-centric paradigm of Wireless sensor networks requires application-aware communication protocols.

Although a considerable amount of recent papers have focused on cross-layer design and improvement of protocols for WSNs, a systematic methodology to accurately model and leverage cross-layer interactions is still missing. With this respect, the design of networking protocols for multi-hop wireless ad hoc and sensor networks can be interpreted as the distributed solution of resource allocation problems at different layers. However, while most of the existing studies decompose the resource allocation problem at different layers, and consider allocation of resources at each layer separately, we review recent literature that has tried to establish sound cross-layer design methodologies based on the joint solution of resource allocation optimization problems at different layers.

Several open research problems arise in the development of systematic techniques for cross-layer design of WSN protocols. In this chapter, we describe the performance improvement and the consequent risks of a cross-layer approach. We review literature proposing precautionary guidelines and principles for cross-layer design, and suggest some possible research directions.

We also present some concerns and precautionary considerations regarding cross-layer design architectures. A cross-layer solution, in fact, generally decreases the level of modularity, which may loosen the decoupling between design and development process, making it more difficult to further design improvements and innovations. Moreover, it increases the risk of instability caused by unintended functional dependencies, which are not easily foreseen in a non-layered architecture.

This chapter is organized as follows. In Section 2, we overview the communication protocols devised for WSNs that focus on cross-layer design techniques. We classify these techniques based on the network layers they aim at replacing in the classical OSI (Open System Interconnection) network stack. In Section 3, a new communication paradigm, i.e., cross-layer module is introduced. In Section 4, we discuss the resource allocation problems that relate to the cross-layer design and the proposed solutions in WSNs. Based on the experience in cross-layering in WSNs, in Section 5 we present the potential open problems that we foresee for WSNs.

2 Pair-Wise Cross Layer Protocols

In this section, we overview significant findings and representative communication protocols that are relevant to the cross-layering philosophy. So far, the term cross-layer has carried at least two meanings. In many papers, the cross-layer interaction is considered, where the traditional layered structure is preserved, while each layer is informed about the conditions of other layers. However, the mechanisms of each layer still stay intact. On the other hand, there is still much to be gained by rethinking the mechanisms of network layers in a unified way so as to provide a single communication

module for efficient communication in WSNs. In this section, we only focus on the pair-wise cross-layer protocols and defer the discussion of cross-layer module design, where functionalities of multiple traditional layers are melted into a functional module, to Section 3.

The experience gained through both analytical studies and experimental work in WSNs revealed important interactions between different layers of the network stack. These interactions are especially important for the design of communication protocols for WSNs. As an example, in (Ganesan 2002), the effect of wireless channel on a simple communication protocol such as flooding is investigated through testbed experiments. Accordingly, the broadcast and asymmetric nature of the wireless channel results in a different performance than that predicted through the unit disk graph model (UGM). More specifically, the asymmetric nature of wireless channels introduces significant variance in the hop count between two nodes. Furthermore, the broadcast nature of the wireless channel results in significantly different floods trees than predicted by the unit disk graph model (Ganesan 2002). Similarly, in (Zuniga 2004), the experimental studies reveal that the perfect-reception-within-range models can be misleading in performance evaluations due to the existence of a transitional region in low power links. The experiment results reported in (Zuniga 2004) and many others show that up to a certain threshold internodes distance, two nodes can communicate with practically no errors. Moreover, nodes that are farther away from this threshold distance are also reachable with a certain probability. While this probability depends on the distance between nodes, it also varies with time due to the randomness in the wireless channel. Hence, protocols designed for WSNs need to capture this effect of low power wireless links. Moreover, in (Shih 2001), guidelines for physical-layer-driven protocol and algorithm design are investigated. These existing studies strongly advocate that communication protocols for WSNs need to be redesigned considering the wireless channel effects. Similarly, as pointed out in (Vuran 2005), the interdependency between local contention and end-to-end congestion is important to be considered during the phase of protocol design. The interdependency between these and other network layers calls for adaptive cross-layer mechanisms in order to achieve efficient data delivery in WSNs.In addition to the wireless channel impact and cross-layer interactions, the content of the information sent by sensor nodes is also important in cross-layer protocol design. In fact, the spatial, temporal, and spatio-temporal correlation is another significant characteristic of WSNs. Dense deployment of sensor nodes results in the sensor observations being highly correlated in the space domain. Similarly, the nature of the energy-radiating physical phenomenon yields temporal correlation between each consecutive observation of a sensor node. Furthermore, the coupled effects of these two sources of correlation results in spatio-temporal correlation. Exploiting the spatial and temporal correlation further improves energy efficiency of communication in WSNs. In (Vuran 2004) and (Vuran 2006-2), the theory of spatial, temporal, and spatio-temporal correlation in WSNs is developed. The correlation between the observations of nodes are modelled by a correlation function based on two different source models, i.e., point and field sources. Based on this theory, the estimation error resulting in

exploiting the correlation in the network can be calculated. This error is defined as distortion. In Figs. x.1 and x.2, the effect of spatial and temporal correlation on the distortion in event reconstruction is shown, respectively. In general, lower distortion results in more accurate estimation of the event features. Hence, using more number of nodes in an event location as shown in Fig. x.1 or sampling the physical locations in higher frequency as shown in Fig. x.2 results in lower distortion. However, Fig. x.1 reveals that, by using a small subset of nodes for reporting an event, e.g., 15 out of 50, the same distortion in event reconstruction can be achieved. Similarly, by reducing the sampling rate of sensor nodes, the same distortion level can be achieved as shown in Fig. x.2 due to correlation between samples. As a result, the redundancy in the sensor readings can be removed. These results reveal that, significant energy savings are possible when the correlation in the content of information is exploited. Moreover, in Fig. x.3, the feasible regions for number of nodes that are reporting an event and their reporting frequency tuple, (M,f), are shown for a given distortion constraint D_{max}. It is clearly shown that, using maximum values for both of these operation parameters may decrease distortion and that these parameters need to be collaboratively selected inside the feasible region using distributed protocols. In the following sections, we will describe two approaches in MAC and transport layers that exploit the spatial correlation in WSNs.

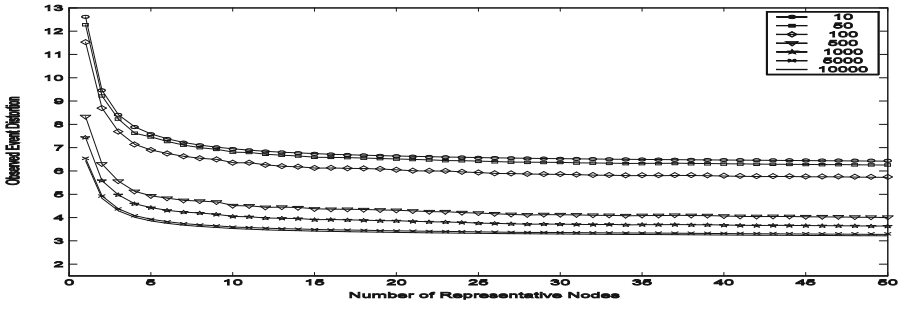


Fig. x.1. Observed event distortion vs. changing number of representative nodes

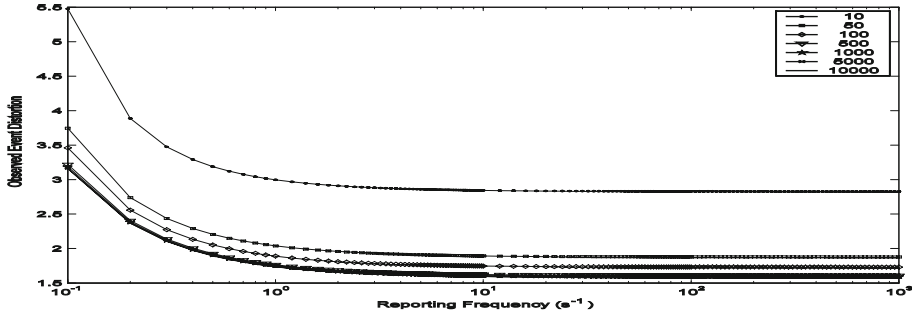


Fig. x.2. Observed event distortion vs. varying normalized reporting frequency (Vuran 2004)

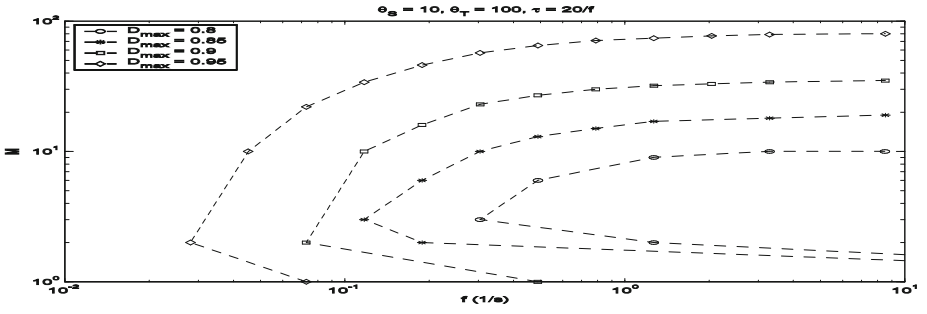


Fig. x3. Number of nodes vs. sampling rate, (M,f) tuples meeting various D_{max} constraints (Vuran 2006-2)

In the following, the literature of WSN protocols with cross-layer principles is surveyed. We classify these studies in terms of interactions or modularity among physical (PHY), medium access control (MAC), routing, and transport layers.

2.1 Transport and PHY Interactions

Transport layer functionalities, such as congestion control and reliability management, depend on the underlying physical properties of both the sensor transceiver and the physical phenomenon that is sensed. More specifically, the transmit power of the sensor nodes directly affects the one-hop reliability. This in effect improves end-to-end reliability. However, increasing the transmit power increases the interference range of a node and may cause increased contention in the wireless medium leading to overall network congestion (Akyildiz 2006). On the other hand, the spatial and the temporal correlation in the content of information enable energy efficient operation by definition of new reliability concepts (Akan 2005). In this section, we overview two representative solutions for pair-wise cross-layer protocols between transport and PHY layers.

In (Chiang 2005), a cross-layer optimization solution for power control and congestion control is considered. More specifically, analysis of interactions between power control and congestion control is provided, and the trade-off between the layered and the cross-layer approach is presented, as further discussed in Section 4. In this analysis, a CDMA-based physical layer is assumed. Consequently, the received signal of a node is modelled as a global and nonlinear function of all the transmit powers of the neighbour nodes. Based on this framework, a cross-layer communication protocol is proposed, where the transmit power and the transmission rate are jointly controlled. The nodes control their transmit power based on the interference of other nodes and determine the transmission rate accordingly. However, the proposed solutions only apply to CDMA-based wireless multihop networks, which may not apply to a large class of WSNs where CDMA technology is not feasible.

The spatial correlation between sensor nodes is exploited in (Akan 2005) with the definition of a new reliability notion. In conventional networks, since the information sent by different entities are independent of each other, a one-to-one and end-to-end reliability notion is used. In WSNs, however, the end user, e.g., sink, is often

interested in the physical phenomenon in the vicinity of a group of sensors instead of the individual readings of each sensor. Consequently, in (Akan 2005), the event-to-sink reliability notion is defined for data traffic from sensors to the sink. This notion relies on the fact that the readings of a group of sensors in an event area are spatially correlated and the reporting rate of these sensors can be collectively controlled to ensure both reliability and prevent congestion. As a result, in event to sink reliable transport (ESRT) protocol, the transmission rate of sensors nodes are controlled by the sink iteratively through calculations during a decision interval.

2.2 MAC and PHY Interactions

As explained above, physical layer properties of WSNs necessitate both channel-aware and physical phenomenon-aware design techniques. This necessity is also valid for medium access control (MAC) protocols. In this section, we present two major approaches in pair-wise interaction for MAC and PHY layers.

The non-uniform properties of signal propagation in low power wireless channels need to be considered in MAC protocol design. MAC protocols aim at providing collision-free access to the wireless medium, and this collision can only be prevented by accurate knowledge of potential interfering nodes. Hence, an accurate wireless channel model is required for both evaluation and design of MAC protocols. In (Haapola 2005), the energy consumption analysis for physical and MAC layers is performed for MAC protocols. In this analysis, the energy consumption due to both processing and transmission is considered. Generally, in ad-hoc networks, multi-hop communication is preferred since transmission power is reduced. However, in WSNs, where processing and communication energy consumption are comparable, this preference is not that clear. Especially for low duty cycle networks, energy consumption due to processing may become comparable to energy consumption due to communication. In this analysis, this trade-off is investigated and it is concluded that single-hop communication can be more efficient when real radio models are used. This result necessitates new techniques for MAC protocols since the number of potential interferers increases significantly when single-hop communication is considered. Although this is an interesting result, the analysis in (Haapola 2005) is based on a linear network and it is necessary to generalize this result to networks with arbitrary topologies.

In addition to the characteristics of the wireless channel and the radio circuitry, the content of the information that will be sent by sensor nodes is also important in MAC design. The content of this information is closely related to the physical properties of the physical phenomenon since WSNs are primarily developed for sensing this phenomenon in the environment. As shown in Fig. x.1, the spatial correlation between the information each sensor node gathers can be exploited for energy efficient operation. Furthermore, since the MAC layer coordinates interactions between closely located nodes, this layer is a perfect fit for exploiting spatial correlation. Consequently, a cross-layer solution among MAC layer, physical phenomenon, and the application layer for WSNs is proposed in (Vuran 2006). The main motivation behind this solution is illustrated in Fig. x.4. Due to the spatial correlation between closely located nodes, in WSNs, a node may contain highly correlated sensor readings as its neighbours. Hence, any information sent by these neighbours may be redundant

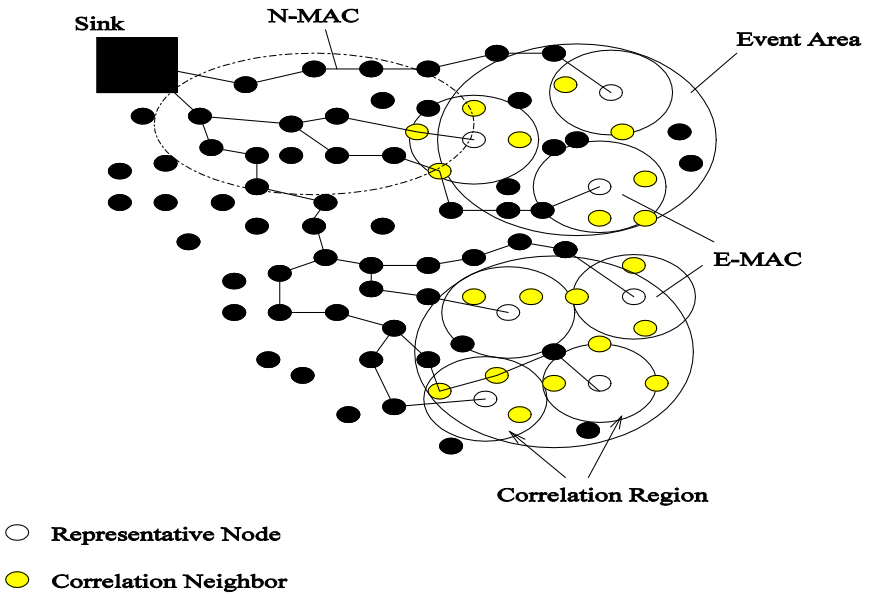


Fig. x.4. CC-MAC protocol and its components Event-MAC (E-MAC) and Network-MAC (N-MAC). The representative node transmits its record on behalf of the entire correlation region, while all correlation neighbours suppress their transmissions (Vuran 2006)

once this node sends its information. Based on the rate-distortion theory, in (Vuran 2006), it is shown that a sensor node can act as a representative node for several other sensor nodes as shown in Fig. x.4. Accordingly, a distributed, spatial correlation-based collaborative medium access control (CC-MAC) protocol is proposed. In this protocol, using the statistical properties of the WSN topology, the maximum distance, d_{corr}, at which two nodes are still highly correlated, given a distortion requirement, is calculated at the sink. Each node then contends for the medium only if it does not hear any node in its correlation region transmitting information. This operation constructs correlation clusters as shown in Fig. x.4. As a result, lower number of communication attempts are performed, which leads to lower contention, energy consumption, and latency while achieving acceptable distortion for reconstruction of event information at the sink. Simulation results in (Vuran 2006) show that this cross-layer interaction results in high performance in terms of energy, packet drop rate, and latency compared to MAC layer protocols designed for WSNs.

3 Cross-Layer Module Design

3.1 Related Work

In addition to the proposed protocols that focus on pair-wise cross-layer interaction, more general cross-layer approaches among three protocol layers exist. In (Madan 2005),

the optimization of transmission power, transmission rate, and link schedule for TDMA-based WSNs is proposed. The optimization is performed to maximize the network lifetime, instead of minimizing the total average power consumption. In (Cui 2005), joint routing, MAC, and link layer optimization is proposed. The authors consider a variable-length TDMA scheme and MQAM modulation. The optimization problem considers energy consumption that includes both transmission energy and circuit processing energy. Based on this analysis, it is shown that single-hop communication may be optimal in some cases where the circuit energy dominates the energy consumption instead of transmission energy. Although the optimization problems presented in this work are insightful, no communication protocol for practical implementation is proposed. Moreover, the transport layer issues such as congestion and flow control are not considered.

A cross-layer approach, which considers routing, MAC, and PHY layers, is also proposed in (Kuruvila 2005). In this work, a MAC protocol is proposed such that the number of acknowledgements sent to the sender depends on the packet reception probability of the node. Moreover, the optimum hop distance to minimize the hop count is found to be less than the transmission range of a node, i.e., $0.72 \cdot R$, which motivates that nodes at the boundary of the transmission range should not be chosen as next hop. Finally, various combinations of greedy and progress-based routing algorithms are simulated showing the advantages of this cross-layer approach over well-known layered protocols.

Although the existing solutions incorporate cross-layer interactions into protocol design, the layering concept still remains intact in these protocols. However, there is still much to be gained by rethinking the functionalities of each protocol layer and melting them into a single cross-layer module. In the following section, we overview our solution for cross-layer design in WSNs, which incorporates transport, routing, MAC, and physical layer functionalities into a single cross-layer module.

3.2 XLM: Cross-Layer Module

The cross-layer approach emerged recently still necessitates a unified cross-layer communication protocol for efficient and reliable event communication that considers transport, routing, and medium access functionalities with physical layer (wireless channel) effects for WSNs. Here, we overview a new communication paradigm, i.e., cross-layer module (XLM) for WSNs (Akyildiz 2006). XLM replaces the entire traditional layered protocol architecture that has been used so far in WSNs.

The basis of communication in XLM is built on the *initiative* concept. The initiative concept constitutes the core of XLM and implicitly incorporates the intrinsic functionalities required for successful communication in WSN. A node initiates transmission by broadcasting an RTS packet to indicate its neighbours that it has a packet to send. Upon receiving an RTS packet, each neighbour of a node decides to participate in the communication through *initiative determination*. Denoting the initiative as I, it is determined as follows:

$$I = \begin{cases} 1, \ if \ \begin{cases} \xi_{RTS} \geq \xi_{Th} \\ \lambda_{relay} \leq \lambda_{relay}^{Th} \\ \beta \leq \beta^{\max} \\ E_{rem} \geq E_{rem}^{\min} \end{cases} \\ 0, \ otherwise \end{cases} \quad \text{(eq. x.1)}$$

where ξ_{RTS} is the received SNR value of the RTS packet, λ_{relay} is the rate of packets that are relayed by a node, β is the buffer occupancy of the node, and E_{rem} is the residual energy of the node, while the terms on the right side of the inequalities indicate the associated threshold values for these parameters, respectively. The initiative I is set to 1 if all four conditions in (x.1) are satisfied. The first condition ensures that reliable links be constructed for communication. The second and third conditions are used for local congestion control in XLM. The second condition prevents congestion by limiting the traffic a node can relay. The third condition ensures that the node does not experience any buffer overflow. The last condition ensures that the remaining energy of a node E_{rem} stays above a minimum value $E_{rem}^{\min}$.

The cross-layer functionalities of XLM lie in these constraints that define the initiative of a node to participate in communication. Using the initiative concept, XLM performs local congestion control, hop-by-hop reliability, and distributed operation. For a successful communication, a node first initiates transmission by broadcasting an RTS packet, which serves as a link-quality indicator and also helps the potential destinations to perform receiver-based contention. Then, the nodes that hear this initiation perform initiative determination according to (x.1). The nodes that decide to participate in the communication contend for routing of the packet by transmitting CTS packets. The waiting time for the CTS packet transmission is determined based on the advancement of a node for routing (Akyildiz 2006). Moreover, the local congestion control component of XLM ensures energy efficient as well as reliable communication by a two-step congestion control. Analytical performance evaluation and simulation experiment results show that XLM significantly improves the communication performance and outperforms the traditional layered protocol architectures in terms of both network performance and implementation complexity.

3.3 Cross-Layer Resource Allocation

Although a considerable amount of recent papers have focused on cross-layer design and improvement of protocols for WSNs, a systematic methodology to accurately model and leverage cross-layer interactions is still largely missing. With this respect, the design of networking protocols for multi-hop wireless ad hoc and sensor networks can be interpreted as the (possibly distributed) solution of resource allocation

problems at different layers. From an engineering perspective, most networking problem can in fact be seen as resource allocation problem, where users (network nodes) are assigned resources (power, time slots, paths, rates, etc.) under some specified system constraints. Resource allocation in the context of multi-hop wireless networks has been extensively studied in the last few years, typically with the objectives of maximizing the network lifetime (Chang 2000), minimizing the energy consumption (Melodia 2005), or maximizing the network throughput (Jain 2003). However, most of the existing studies decompose the resource allocation problem at different layers, and consider allocation of the resources at each layer separately. In most cases, resource allocation problems are treated either heuristically, or without considering cross-layer interdependencies, or by considering pair-wise interactions between isolated pairs of layers.

A typical example of the tight coupling between functionalities handled at different layers is the interaction between the congestion control and power control mechanisms (Chiang 2005). The congestion control regulates the allowed source rates so that the total traffic load on any link does not exceed the available capacity. In typical congestion control problems, the capacity of each link is assumed to be fixed and predetermined. However, in multi-hop wireless networks, the attainable capacity of each wireless link depends on the interference levels, which in turn depend on the power control policy. Hence, congestion control and power control are inherently coupled and should not be treated separately when efficient solutions are sought.

Furthermore, the physical, medium access control (MAC), and routing layers together impact the contention for network resources. The physical layer has a direct impact on multiple access of nodes in wireless channels by affecting the interference at the receivers. The MAC layer determines the bandwidth allocated to each transmitter, which naturally affects the performance of the physical layer in terms of successfully detecting the desired signals. On the other hand, as a result of transmission schedules, high packet delays and/or low bandwidth can occur, forcing the routing layer to change its route decisions. Different routing decisions alter the set of links to be scheduled, and thereby influence the performance of the MAC layer.

Several papers in the literature focus on the joint power control and MAC problem and/or power control and routing issues, although most of them study the interactions among different layers under restricted assumptions. In Section 1, we report a set of meaningful examples of papers considering pair-wise resource allocation problems. In particular, we report examples of joint scheduling and power control, joint routing and power control, and joint routing and scheduling. In Section 2, we describe previous work that dealt with cross-layer optimal resource allocation at the physical, MAC, and routing layer. In Section 3, we discuss recent work on cross-layer design techniques developed within the framework of network utility maximization. Since these techniques often naturally lead to decompositions of the given problem and to distributed implementation, these can be considered promising results towards the development of systematic techniques for cross-layer design of sensor networks. Most of the papers described in this section consider general models of multi-hop wireless networks, and try to derive general methodologies for cross-layer design of wireless networks. Hence, unless otherwise specified, the techniques described here equally apply to the design of sensor networks and general purpose ad hoc networks.

4 Open Research Problems

As explained in Sections 2, 3 and 4, there exists remarkable effort on cross-layer design in order to develop new communication protocols. However, there is still much to be gained by rethinking the protocol functions of network layers in a unified way so as to provide a single communication module that limits the duplication of functions, which often characterizes a layered design, and achieves global design objectives of sensor networks, such as minimal energy consumption and maximum network lifetime. In fact, research on cross-layer design and engineering is interdisciplinary in nature and it involves several research areas such as adaptive coding and modulation, channel modelling, traffic modelling, queuing theory, network protocol design, and optimization techniques.

There are several open research problems toward the development of systematic techniques for cross-layer design of wireless sensor network protocols. It is needed to acquire an improved understanding of energy consumption in WSNs. In fact, existing studies on cross-layer optimization are mostly focused on jointly optimizing functionalities at different layers, usually with the overall objective of maximizing the network throughput. Conversely, in WSNs the ultimate objective is usually to minimize the energy consumption and/or to maximize the network lifetime. Hence, further study is needed to develop models and methodologies suitable to solve energy-oriented problems.

It is also necessary to develop sound models to include an accurate description of the end-to-end delay in the above framework as results from the interaction of the different layers. In particular, there is a need to develop mathematical models to accurately describe contention at the MAC layer. This would allow determining the set of feasible concurrent transmissions under different MAC strategies. This is particularly important for the design of sensor network protocols for monitoring applications that require real-time delivery of event data, such as those encountered in wireless sensor and actor networks (WSAN) (Akyildiz 2004).

Moreover, characteristics of the physical layer communication, such as modulation and error control, that impact the overall resource allocation problem should be incorporated in the cross-layer design. For example, in future wireless communications, adaptive modulation could be applied to achieve better spectrum utilization. To combat different levels of channel errors, adaptive forward error coding (FEC) is widely used in wireless transceivers. Further, joint consideration of adaptive modulation, adaptive FEC, and scheduling would provide each user with the ability to adjust the transmission rate and achieve the desired error protection level, thus facilitating the adaptation to various channel conditions (Liu 2005)(Cui 2005-2).

Another important open research issue is to study the network connectivity with realistic physical layer. Connectivity in wireless networks has been previously studied (Gupta 1998)(Bettstetter 2002), i.e., stochastic models have been developed to determine conditions under which a network is connected. These results, however, cannot be straightforwardly used, as they are based on the so-called unit disk graph communication model. Recent experimental studies, however, have demonstrated that the effects of the impairments of the wireless channel on higher-layer protocols are not negligible. In fact, the availability of links fluctuates because of channel fading phenomena that affect the wireless transmission medium. Furthermore, mobility of

nodes is not considered. In fact, due to node mobility and node join and leave events, the network may be subject to frequent topological reconfigurations. Thus, links are continuously established and broken. For the above reasons, new analytical models are required to determine connectivity conditions that incorporate mobility and fading channels.

Last but not least, new cross-layer network simulators need to be developed. Current discrete-event network simulators such as OPNET, ns-2, J-Sim, GloMoSim may be unsuitable to implement a cross-layer solution, since their inner structure is based on a layered architecture, and each implemented functionality run by the simulator engine is tightly tied to this architecture. Hence, implementing a cross-layer solution in one of these simulators may turn into a non-trivial task. For this reason, there is a need to develop new software simulators that are based on a new developing paradigm so as to ease the development and test of cross-layer algorithmic and protocol solutions.

5 Conclusions

In this chapter, we reviewed and classified literature on cross-layer protocols, improvements, and design methodologies for wireless sensor networks (WSNs). We overviewed the communication protocols devised for WSNs that focus on cross-layer design techniques. We classified these techniques based on the network layers they aim at replacing in the classical OSI network stack. Furthermore, we discussed systematic methodologies for the design of cross-layer solution for sensor networks as resource allocation problems in the framework of non-linear optimization. We outlined open research issues in the development of cross-layer methodologies for sensor networks and discussed possible research directions.

A cross-layer design methodology for energy-constrained wireless sensor networks is an appealing approach as long as cross-layer interactions are thoroughly studied and controlled. As pointed out in this chapter, in fact, no cross-layer dependency should be left unintended, since this may lead to poor performance of the entire system.

References

[1] Akan, O.B., Akyildiz, I.F.: Event-to-sink reliable transport in wireless sensor networks. IEEE/ACM Transactions on Networking 13(5), 1003–1017 (2005)

[2] Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. Computer Networks (Elsevier) Journal 38(4), 393–422 (2002)

[3] Bettstetter, C.: On the minimum node degree and connectivity of a wireless multihop network. In: Proc. ACM MobiHoc 2002, Lausanne, Switzerland, pp. 80–91 (2002)

[4] Boyd, S., Vandenberghe, L.: Convex optimization (2004); Chang, J.H., Tassiulas, L.: Energy Conserving Routing in Wireless ad Hoc Networks. In: Proc. IEEE Infocom 2000 pp. 22–31. Cambridge University Press (2000)

[5] Chiang, M.: Balancing transport and physical layers in wireless multihop networks: Jointly optimal congestion control and power control. IEEE Journal of Selected areas in Communications 23(1), 104–116 (2005)

[6] Kozat, U.C., Koutsopoulos, I., Tassiulas, L.: A framework for cross-layer design of energy-efficient Communication with QoS provisioning in multi-hop wireless networks. In: Proc. INFOCOM 2004, Honk Kong, pp. 1446–1456 (2004)
[7] Uruvila, J., Nayak, A., Stojmenovic, I.: Hop count optimal position based packet routing algorithms for ad hoc wireless networks with a realistic physical layer. IEEE Journal of Selected Areas in Communications 23(6), 1267–1275 (2005)
[8] Lee, J.W., Chiang, M., Calderbank, R.: Price-based distributed algorithms for optimal rate-reliability trade-off in network utility maximization. To appear in IEEE Journal of Selected Areas in Communications (2006)
[9] Li, Y., Ephremides, A.: Joint scheduling, power control, and routing algorithm for ad-hoc wireless networks. In: Proc. Hawaii International Conference on System Sciences, Big Island, HI, USA (2005)
[10] Liu, Q., Zhou, S., Giannakis, G.B.: Cross-layer scheduling with prescribed QoS guarantees in adaptive wireless networks. Journal of Selected Areas in Communications 23(5), 1056–1066 (2005)

Semantic Based Category-Keywords List Enrichment for Document Classification

Upasana Pandey, S. Chakraverty, Richa Mihani, Ruchika Arya,
Sonali Rathee, and Richa K. Sharma

Netaji Subhas Institute of Technology (NSIT), Azad Hind Fauj Marg, Sector-3, Dwarka,
New Delhi-110078
coe.upasana@gmail.com, apmahs@rediffmail.com,
richamihani9@gmail.com, ruchiarya31@gmail.com,
sonali.rathee@gmail.com, enigmaticricha@gmail.com

Abstract. In this paper we present a text categorization technique that extracts semantic features of documents to generate a compact set of keywords and uses the information obtained from those keywords to perform text classification. The algorithm reduces the dimensionality of the document representation using overlapping semantics. Later, a keyword-category relationship matrix computes the extent of membership of the documents for various input predefined categories. The category of the document is then derived from membership metrics. Also, Wikipedia is used for the purpose of category lists enrichment. The proposed work has shown a new direction towards document classification for web applications.

Keywords: Overlapping Semantics, Lexical Chaining, Membership metrics.

1 Introduction

The modern information age produces vast amounts of textual data, which can be considered largely as unstructured data. If this data is properly organized and classified, then retrieving the relevant information from the maze of data becomes much simpler. With the exponential growth of documents, the need for automated methods to organize and classify the documents in a reliable manner becomes inevitable.

Text Categorization (TC), also known as Text Classification, is the process of dividing a set of text documents into specified categories [1]. TC involves a series of challenging tasks such as: (i) setting up and pre-processing test-set data and training-set data, (ii) organizing the knowledge-base of text in some form such as ontology or a database with meta-data, (iii) feature selection and extraction,(iv) designing a classifier and (v) testing and evaluating the classifier. Manual classification is not only time consuming but more vulnerable to human errors as well. Besides, domain experts in the areas of pre-defined categories would be needed.

Over the past decade or so, researchers have tried to automate each of the above steps using a plethora of techniques. Traditionally, TC requires a substantial amount of manually labeled documents for classification which is often impractical in real-life settings. Keyword-based TC methods aim at a more practical setting. Each category is

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 297–309.
springerlink.com

represented by a list of characteristic keywords which capture the category meaning. The effort is then reduced to providing an appropriate keyword list per category, a process that can be automated.

Classification is then achieved by measuring similarity between the pre-defined category names and their keywords and the documents to be classified.

For the rest of the paper, in section 2 we present a background of the field and the relevance of context based TC. Section 3 presents our proposed algorithm. Section 4 presents experiment and discussion. In section 5, we compare the proposed TC scheme with current approaches and overview its implications and advantages. We conclude in Section 6.

2 Background and Motivation

Document classification broadly follows two approaches: Supervised document classification where some external mechanism provides information to guide the correct classification, and unsupervised document classification where the classification must be done without reference to external information. Semi-supervised document classification also exists, whereby parts of the documents are labeled by an external mechanism [11].

TC methods can further be classified as Statistical and Semantic methods. Statistical techniques consider words of a document as unordered or independent elements [12] and simply compute the frequency of the feature items. They do not take into account the characteristics of position and ignore the fact that words at different positions have different contributions to the theme of the article. The Term frequency-Inverse document frequency (Tf-Idf) weight is often used in information retrieval and text mining. This statistical measure evaluates the importance of a word in a document [13]. Semantic methods exploit the information provided by the word context. These methods exploit the relationship among the words of a document in order to evaluate their semantic relevance to a given category [12].

Statistical approaches including support vector machines [2], memory-based learning [3], rough set [4], neural networks [5], Bayesian classifiers [6] and sparse binary polynomial hash [7] have reported significant successes in TC and web applications. Their applicability in further improving the quality of TC seems to have reached a pinnacle. On the other hand, semantic feature extraction offers a lot of scope to experiment on the relevancy among words in a document and exploit it to achieve improved results. Context can be interpreted in a variety of ways like lexical cohesiveness [14], use of lexical units [24], syntactical constructs such as Parts of Speech [16], semantic constructs [16] and distance based methods [12]. Implicit correlation between words is expressed through Latent Semantic Analysis (LSA) [15]. With such as wide spectrum, different context-oriented features can be flexibly combined together to target different applications.

For text categorization using the context-oriented approach, a central concern is to automatically cull out a set of training documents from given corpus which can be associated with a certain category. Automatic generation of keyword-list per category is a potential area that motivates the use of contextual information to find suitable keywords for each category.

Lexical cohesion is a type of textual cohesion that allows the use of similar meaning words through synonyms, generalization of concepts through the use of hypernyms, specialized versions of a concept through the use of hyponyms or enunciates parts of an object through meronyms. In [16], the authors demonstrated how well organized background knowledge in form of simple ontologies can improve text classification results. Although designed primarily as a lexical database, WordNet can be used as ontology [17, 18, 20]. For each concept present in a document, the referring terms can be found in WordNet starting from the relevant senses of the category name and transitively following relation types that correspond to lexical references. Thus, the concern about extracting a suitable list of keywords for a given category can be addressed by utilizing the WordNet.

In addition to the keywords obtained from WordNet, Wikipedia [28] can also be used to generate more exhaustive keyword lists per category. Wikipedia is a free, open content online encyclopedia created through the collaborative effort of a community of users known as Wikipedians. Like an encyclopedia, there are an unlimited number of subjects and topics. Wikipedia articles are densely structured. Its articles follow a chain of linkage with of millions of other articles [25].

The motivation of using Wikipedia is its ability to quantify semantic relatedness of texts. Wikipedia handles many fundamental tasks in computational linguistics, including word sense disambiguation, information retrieval, word and text clustering, and error correction [26]. The basic purpose to include Wikipedia is to enhance the keyword list per category which is obtained from WordNet. This way terms which are practically related to the categories are also added. This can be elaborated with the help of the fact that initially after using WordNet, the category list for space did not have the name of the planets. So, it was realised that Wikipedia can be used as a machine learning tool [27]. As a result, the keywords which are obtained from the already classified documents and are closely related to the category are added to the respective category lists. As a result of this, the names of the planets were added in the category list for Space.

Our impression of the real world is based on relative concepts, which do not have specifically defined boundaries. Concepts like few, many, small, tall, much smaller than, etc. are relative and depend on the sense they are used in. That is to say that they are true only to some degree, and are false to some extent as well. These concepts can be called fuzzy or vague concepts. The way human brain works with them, a computer does not work the same way since it works on binary logic (strings of 0s and 1s).

Since, there is no clear separation between two or more categories and fuzzy logic is a good way to deal with such fuzzy boundaries. In contrast to binary logic which defines crisp boundaries, fuzzy logic deals with the extent of relevance. The algorithm is motivated from fuzzy logic though no fuzzy rules are incorporated in the algorithm.

In this paper, we work on semantic information obtained from the WordNet [18].This information is used to determine the contextual relationships between the words of a document so as to generate a set of keywords for each category. The set of keywords is then compacted using the concept of overlapping semantics in order to reduce the dimensionality of document representation. The context-driven statistical information is then fed into a fuzzy model to improve the efficiency of TC. At last we use the Wikipedia [28] for category-keyword list enrichment for better classification.

In concordance with the generic model for TC, our algorithm uses automatically pre-classified training documents for each category and classifies unseen documents.

3 Proposed Scheme for TC

The task of classifying the test documents is divided into two parts:-

A. Keyword extraction: The first part exploits the semantic features encapsulated in the documents. Category names are the first input to the system. A set of keywords that possess strong semantic correlation with the category name is extracted. The WordNet lexical database is utilized as ontology for this purpose. Concepts in each document are expanded by navigating through word sense nodes.

Wikipedia is also used to make the keyword lists more exhaustive by adding first-level hyperlinks (related to the respective category names) to the category lists obtained from WordNet. This step is referred to as Wikipedia-pre-processing for the purpose of category-lists enrichment. Each category is thus represented by a list of characteristic keywords which capture the category meaning. Next, we extract those keywords that represent the document¡¦s context greatly. This is done with the help of overlapping semantics given in [19]. Overlapping semantics helps in reducing the dimensionality of the document. The fundamental premise is that if two features in a document have a common synonym set, then the corresponding features represent the document¡¦s meaning greatly.

B. Membership metrics: After acquiring the keyword-list per category and reducing the features of the document, we construct a keyword-category relationship matrix. This matrix is used to compute the membership of the document in each of the predefined categories. The document is classified into that category for which the membership has the maximum value.

Few assumptions are made before implementing the algorithm. Firstly, the categories are previously identified for which the keywords are populated in the first step of algorithm. Secondly, there is no explicit set of training and testing documents chosen. The algorithm is derived with the aim of self-enrichment of keyword list, per category, with each classifying document. The pseudo code for the proposed algorithm is given in figure 1 and explained below.

Step 1: Acquiring keyword-list per category: Words that bear a strong semantic connection and lexical reference with category name are keywords. Words that are lexically related to a category are collected from the WordNet source. First synonyms were extracted. Then hypernyms of each term are transitively located to allow generalisation. Hypernyms are collected up to the first level only to avoid complexity. Next hyponyms at immediate lower level are located. Meronyms and coordinate terms are also considered to get an ontological list of each category. Then, Wikipedia is pinged with the category name. The list of first-level hyperlinks is created and the words which were absent initially were added to the category list. At the end of this process, the system is armed with an initial set of keywords that explicitly represent the meaning of the category.

Step 2: Pre-processing document: The document to be classified, say Di, needs to be pre-processed first. This includes:
2.1 Stop-word removal: All the stop words, i.e. words that appear frequently but do not affect the context are removed from the document. Examples of such words include a, an, and, the, that, it, he, she etc.
2.2 Tokenizing: The document is fragmented into a set of tokens separated by some delimiters, e.g. whitespaces. These tokens (or terms) can represent words, phrases or any keyword patterns.
2.3 Stemming: The resulting set of tokens is replaced by their base form to avoid treating different forms of the same word as different attributes. This reduces the size of the attribute set. For example, both celebration and celebrating are converted to the same base form celebrate.
2.4 Term Weighting: For each token w_i in the document's token set, its frequency f_i is computed.
Step 3: Reducing dimensionality by overlapping semantics: A word possibly has many meanings. WordNet expresses a meaning with a semantic set. If two features/tokens in a document have a common semantic set, then such features can be considered to represent the meaning of the document greatly. The algorithm works as follows:
3.1 Take all the tokens in the document as the set W.
3.2 Take the set of all semantics of W as S using WordNet.
3.3 Indicate semantic subset SubS that appeared repeatedly in S.
3.4 Get the words corresponding to semantics subset SubS from the set W.

The final tokens thus collected from the base set signify the document's meaning to a greater extent and reduce the dimensionality of document representation. Let N_k be the number of tokens representing the document after overlapping semantics.
Step 4: Generating keyword-category relationship matrix: We now construct the category-keyword relationship matrix. In this matrix number of rows is equal to the number of tokens in the document which have been derived by using overlapping semantics and number of columns is equal to number of categories. The elements of this matrix denote the membership of a token in a category. The membership $R_{i,k}$ of a token w_i to a given category C_k is obtained by $P_i(C_k)$; the presence (=1) or absence (=0) of a token w_i in the category C_k divided by the presence or absence of the same token in all the pre-defined categories. The presence or absence of the token is checked in the ontological lists of the categories.
Step 5: Computation of membership metric: For each column in the matrix, a category membership metric μ_k, that reflects the degree of membership of a document D_i to a given category C_k, is computed using the frequencies of the tokens in the document and their corresponding membership values. This metric, given in equation 1, is computed for all the columns, i.e. for the predefined categories.
Step 6: Assigning document to category: The membership metrics obtained for all the columns are compared. The document is classified to that category for which this metric has the maximum value.
Step 7: Wikipedia Training: Let us say the document D_i is classified to category C_{di}. Thus, now for all those tokens in set N_k of document D_i which are absent in category C_{di} keyword list (or have membership metric of zero) are considered for further

enrichment of the C_{di} category keyword list. Let us call a set of such tokens as E_{di}. For each token in E_{di}, the first-level hyperlinks list is obtained using Wikipedia. This list is then intersected with the C_{di} category keyword list. If the intersection results in number of common words greater than a specific threshold, then those tokens are added to the category list as a keyword, resulting in enriched category set. In the implementation, the threshold is set as 5 common words. Once an enrichment of category list is completed for a document D_i, the process for another document D_k is started from the step1 of the algorithm for same set of predefined categories.

So, for each new test document, classifier starts from step 1 and if again find some new tokens of document with membership equal to zero, it starts wikipedia training for these new tokens of document to enrich category-keywords list.

4 Experiment and Discussion

All steps of the proposed TC algorithm were automated on the Windows OS platform using server side scripting language PHP. The algorithm was tested on a set of four predefined categories namely: Fuel, Computers, Sports and Space; and 4 files taken from the standard corpora 20Newsgroups, each belonging to one of the categories specifically, namely: fuelfile, spacefile, compfile and sportsfile. The results so obtained are summarized in table-1, which explicitly mentions the final membership values of each file in respective categories.

Table 1. Membership metric for each category

Category	Fuel	Computer	Sports	Space
Fuelfile	0.280	0.056	0.000	0.000
Spacefile	0.000	0.055	0.058	0.100
Compfile	0.000	0.330	0.069	0.020
Sportsfile	0.000	0.000	0.244	0.000

We illustrate the detailed working of our algorithm on a document, Fuelfile. We initially list results without the incorporation of Wikipedia as a training module and later signify the strong impact of Wikipedia use as a tool to enhance the category list.

This document was pre-processed as per the proposed steps. After the preprocessing the size of the file reduced from 900 bytes to 291 bytes resulting in a compression of 67.67%.The meaning sets of tokens were retrieved from WordNet source. The application of the concept of overlapping semantics on these meaning sets resulted in a final document of reduced dimensionality with only seven words representing the meaning of the document greatly. The achieved compression was 81.44% from 291 bytes to 54 bytes.

The keyword-category relationship matrix thus generated has 7 rows corresponding to the 7 tokens extracted and the ontological list of 4 categories represented by 4 columns. The presence or absence of each token in the given category was checked from the ontological list of the categories. Table 2 shows the presence/absence matrix thus derived.

Table 2. Tokens presence/absence - Category matrix

Categories/Tokens	Fuel	Computers	Sports	Space
Distillate	1	1	0	0
Fuel	1	0	0	0
Stocks	0	0	0	0
Cargo	0	0	0	0
April	0	0	0	0
Explosion	0	0	0	0
Gasoline	1	0	0	0

The membership values derived from the presence/absence matrix are tabulated in Table 3.

Table 3. Tokens membership values

Categories/Tokens	Fuel	Computers	Sports	Space
Distillate	0.5	0.5	0	0
Fuel	1	0	0	0
Stocks	0	0	0	0
Cargo	0	0	0	0
April	0	0	0	0
Explosion	0	0	0	0
Gasoline	1	0	0	0

Finally, the membership metrics showing the extent to which the test document belongs to each category were calculated by the formula:

$$\mu_k = \Sigma^{N}{}_{k} f_i \, R_i(C_k) / \Sigma f_i$$

It was found that the given document had the maximum membership metric for the category Fuel and was thus ascribed to this category.

Now Wikipedia was used to quantify semantic relatedness of texts. Its use as a tool to enrich the respective category lists shown improved results as the degree of belongingness of document increased from 0.28 to 0.644 as shown in Table 4:

Table 4. Membership metric when Wikipedia was used

Category	Fuel	Computer	Sports	Space
Fuelfile	0.644	0.056	0.000	0.000

Although the document was categorized, it was necessary to train the semantic feature extractor to enhance the capabilities of the system. The classified document had tokens not belonging to the respective category but having a high potential to represent the category set. Henceforth, it became necessary to train the system accordingly.

This machine learning step was performed through the resourceful use of Wikipedia. Figure 1 explains the inclusion of Wikipedia in the proposed model.

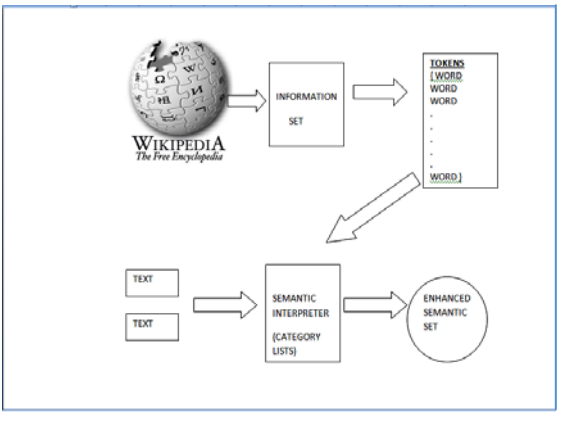


Fig. 1. Wikipedia training

Wikipedia knowledge base is pinged for the information with potential to represent the category list. The information obtained is then tokenized. The intersection of the tokenized information retrieved from Wikipedia and the ontological senses of the words in category list is carried out. The result of this intersection defines the basis for addition of the tokens in the category list. Figure 2 depicts the enhancement in degree of belongingness of the document as experiment was carried out after successive addition of keywords.

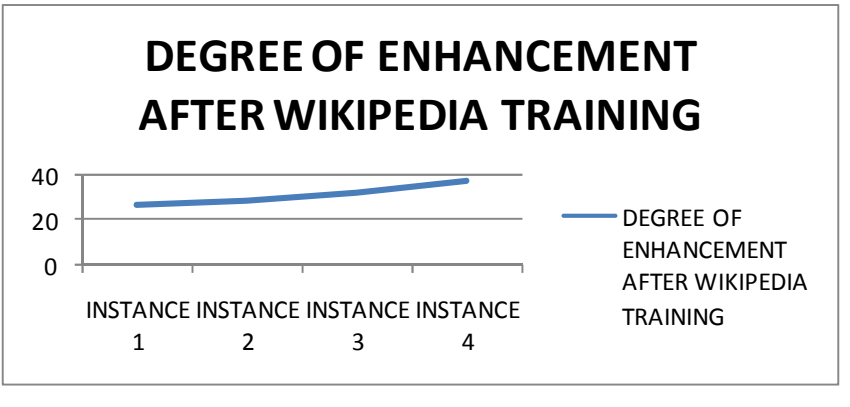


Fig. 2. Degree of enhancement for the document after Wikipedia Training
NOTE: Instances indicate keywords which were added to category lists.

Following the same steps as described above, experiment was first performed on a set of 40 documents from standard corpora, 20Newsgroups [29] and Reuters 21578 [30], where only WordNet was used to initialize the category lists. 10 documents from each category were tested and the results obtained are summarized in Table 5 and graphically represented in Figure 3.

Table 5. Cumulative results on a set of 40 documents using only WordNet to obtain category lists

Category	No of Documents	Correctly Classified	Incorrectly Classified	Accuracy
Computer	10	06	04	60%
Fuel	10	10	00	100%
Space	10	09	01	90%
Sports	10	04	06	40%
Total	40	29	11	72.5%

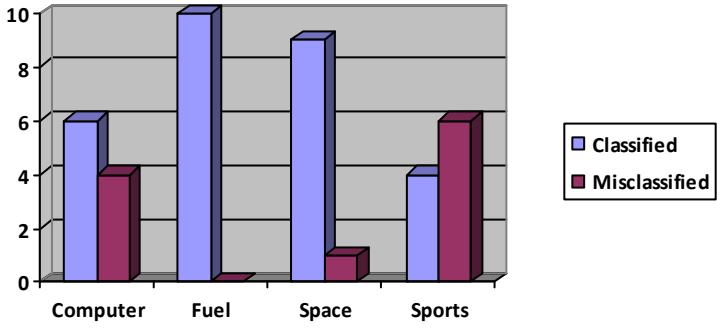


Fig. 3. Results on a set of 40 documents using only WordNet to obtain category lists

The classification accuracy is achieved about 72.5%. Later the experiment was performed on 400 new documents belonging to same standard corpora, 20Newsgroups and Reuters21578. For each category, 100 different documents are taken and using the proposed algorithm, with Wikipedia and WordNet, to initialize the category lists, their membership values are obtained.

The cumulative result thus inferred from the experiment is compiled in Table 6 and graphically represented in Figure 4 in terms of percentage achieved.

The overall average text classification accuracy achieved is about 85.75% using WordNet and Wikipedia to initialize category lists on a set of 400 documents.

Table 6. Cumulative results on a set of 400 documents using WordNet and Wikipedia to initialize category lists

Category	No of Documents	Correctly Classified	Incorrectly Classified	Accuracy
Computer	100	85	15	85%
Fuel	100	87	13	87%
Space	100	82	18	82%
Sports	100	89	11	89%
Total	400	343	57	85.75%

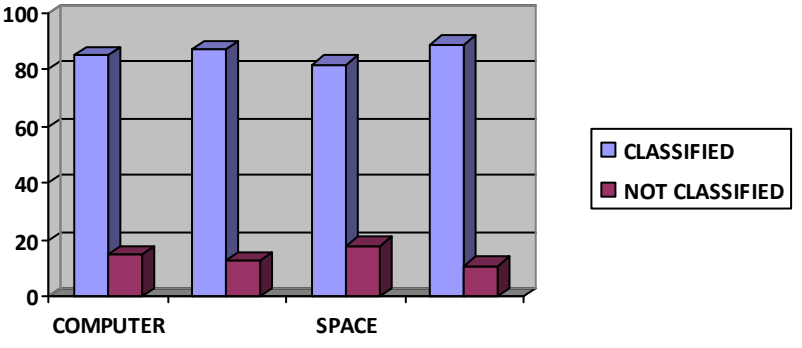


Fig. 4. Percentage of documents correctly classified and misclassified in respective category

So we can conclude here that after including Wikipedia, the Category-Keyword list has been enriched. That may improve the document classification accuracy from 72.5% to 85.75%, which is a good improvement. We have obtained the significant result in terms of contextual document classification that can give a path to improve text classification for other web applications.

5 Comparison with Prior Techniques

The algorithm proposed for text classification proposed in this paper differs from reported schemes with the following innovations:

1. Most authors rely on manually generated keywords for each category [20, 22]. In the proposed algorithm, we build the representative keyword-list per category using WordNet as ontology. We have intentionally derived a broad spectrum of lexical relationships from the WordNet including synonyms, hypernyms, hyponyms as well as meronyms and coordinate terms. This allows the extraction of a richer subset of keywords that bear a strong semantic relationship with the category name.

2. Keyword-category relationship matrix: The keyword-category matrix constructed by our algorithm is different from the one which is built in LSA since the rows in our case represent a reduced set of ontology-extracted keywords. LSA measures the number of times a particular word occurs in the title for all documents taken together [15]. In doing so, LSA assumes that the same word occurring in two different places denote the same concept which causes problems. In contrast, we first check the presence or absence of each document¡¦s keywords that are extracted by a combination of lexical relationships and then reduced by overlapping semantics. We then use keywords frequencies to calculate membership metric for each category thereby classifying the document. Thus in our approach, we take into account the fact that if two features in one category has a common synonym set, then these features represent this category to a greater extent.
3. The proposed algorithm bridges the gap between the statistical and context based techniques. Using semantic feature extraction, we can reduce the vector space dimension of the document to be classified. The idea behind using the fuzzy logic is that the degree of truth of a statement can range between 0 and 1 and is not constrained to the two truth values of classic propositional logic [8], [21].

6 Conclusion and Future Work

In this work, we have implemented text categorization using semantic features. A list of keywords for each category is automatically extracted by using the WordNet lexical database as ontology to derive the synomyms, hypernyms, hyponymns, meronyms and coordinate terms for concepts. The document is thus represented by a set of tokens that have strong semantic correlation with its category. Further dimensionality reduction is performed by applying the concept of overlapping semantics. The extent of belongingness of the test document to the input categories lies between 0 and 1. The benefit of using derived membership formula is that the obtained result is not restricted to two values: true and false. The document is finally classified to that category for which the membership metric has the maximum value. Our experiments demonstrate that this approach efficiently classifies the given document into its most relevant category.

The classifier classifies the documents with an accuracy of approximate 86%. Training using Wikipedia helped to achieve the increase in accuracy as the category lists get enhanced with each successive document it classifies. However, this training step is semi- automated and therefore it is time-consuming. Moreover, the classifier needs to be improved to be able to classify the document to its most relevant category when the membership values coincide in more than one category.

The time complexity of the overlapping semantics algorithm was observed to be $O(n^4)$ and for classification algorithm was found to be $O(n^2)$.

For future work, we intend to integrate lexical chaining to increase the weight of a word on the basis of its surrounding referring words in a document. Moreover, compared to lexical resources such as WordNet, usage of Wikipedia leverages knowledge bases that are orders of magnitude larger and more comprehensive.

References

[1] Wajeed, M.A., Adilakshmi, T.: Text Classification using Machine learning. A Journal of Theoretical and Applied Information Technology 7(2) (2009)

[2] Wang, Q., Guan, Y., Wang, X.: SVM Based Spam Filter with Active and Online Learning. In: Procs. of the TREC Conference (2006)

[3] Androutsopoulos, I., et al.: Learning to filter spam email: a comparison of a naive Bayes and a memory based approach. In: Procs. of the Workshop Machine Learning and Textual Information Access, 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (2000)

[4] Bao, Y., Asai, D., Du, X., Yamada, K., Ishii, N.: An Effective Rough Set-Based Method for Text Classification. In: Liu, J., Cheung, Y.-m., Yin, H. (eds.) IDEAL 2003. LNCS, vol. 2690, pp. 545–552. Springer, Heidelberg (2003)

[5] Li, C.H., Park, S.C.: Text Categorization Based on Artificial Neural Networks. In: King, I., Wang, J., Chan, L.-W., Wang, D. (eds.) ICONIP 2006, Part III. LNCS, vol. 4234, pp. 302–311. Springer, Heidelberg (2006)

[6] Hovold, J.: Naive Bayes Spam Filtering Using Word Position Based Attributes. In: International Conference of Email and Anti Spam (2005)

[7] Yerazunis, W.S.: PhD, Sparse Binary Polynomial Hashing and the CRM114 Discriminator! In: Proc of MIT Spam Conference (2004), http://www.merl.com/papers/docs/TR2004-091.pdf

[8] http://en.Wikipedia.org/wiki/fuzzylogic

[9] http://en.Wikipedia.org/wiki/Lotfi_Zadeh

[10] El-Alfy, E.-S.M., Al-Qunaieer, F.S.: A Fuzzy Similarity approach for Automated Spam filtering. In: Proc. of the 2008 IEEE/ACS International Conference on Computer Systems and Applications, vol. 00, pp. 544–550 (2008)

[11] Chapelle, O., Scholkopf, B., Zien, A.: Book on Semi-Supervised Learning

[12] Ernandes, M., et al.: An Adaptive context Based algorithm For Term Weighting. In: Proc. of the 20th International Joint Conference on Artificial Intelligence, pp. 2748–2753 (2007)

[13] Hu, J.-Z., Shu, J.-B., Huang, Y.-Y.: Text Feature Extraction based on Extension of Topic Words and Fuzzy Set. In: Proc. of 2008 Intl. Conference on Computer Science and Software Engineering (2008)

[14] Teich, E., et al.: Exploring Lexical Patterns in Text: Lexical Cohesion Analysis with WordNet. In: Proc. of Interdisciplinary Studies on Information Structure, vol. 02, pp. 129–145 (2005)

[15] http://en.Wikipedia.org/wiki/LSA

[16] Bloehdron, S., et al.: Boosting for Text Classification with Semantics Features. In: Proc. of the MSW 2004 Workshop at the 10th ACM SIGKDD Conference on Knowledge, Discovery and Data Mining, pp. 70–87 (August 2004)

[17] http://www.webkb.org/interface/categSearch.html?categField=sports

[18] http://www.WordNet-online.com/

[19] Luo, N., et al.: Using CoTraining and Semantic Feature Extraction for Positive and Unlabeled Text Classification. In: Proc. of International Seminar on Future Information Technology and Management Engineering (2008)

[20] Barak, L., et al.: Text Categorization from Category Name via Lexical Reference. In: Proc. of NAACL HLT 2009: Short papers, pp. 33–36 (June 2009)

[21] Haruechaiyasak, C., Shyu, M.-L., Chen, S.-C.: Web Document Classification Based on Fuzzy Association. In: Proc. of the 26th International Computer Software and Applications Conference on Prolonging Software Life: Development and Redevelopment (2002)
[22] Padmaraju, D., et al: Applying Lexical Semantics to Improve Text Classification, http://web2py.iiit.ac.in/publications/default/download/inproceedings.Pdf.9ecb6867-0fb0-48a5-8020-0310468d3275.pdf
[23] Muztaba Fuad, M., Deb, D., Shahriar Hossain, M.: A Trainable Fuzzy Spam Detection System, http://people.cs.vt.edu/msh/papers/trainable.pdf
[24] Pandey, U., et al.: Context Driven Technique for Document Classification. In: Proc. of ACS (2010)
[25] Wikipedia-based Semantic Interpretation for Natural Language Processing by Shaul Markovitch, Department of Computer Science TechnionIsrael Institute of Technology (2009)
[26] Learning to Link with Wikipedia by David Milne, Department of Computer Science, University of Waikato (2008)
[27] Building Semantic Kernels for Text Classification using Wikipedia by Pu Wang, Department of Computer Science, George Mason University (2007)
[28] http://www.wikipeida.org
[29] http://people.csail.mit.edu/jrennie/20Newsgroups
[30] http://www.daviddlewis.com/resources/testcollections/reuters21578

Selection of Fluid Film Journal Bearing: A Fuzzy Approach

V.K. Dwivedi[1], Satish Chand[2], and K.N. Pandey[2]

[1] Department of Mechanical Engineering GLA, University, Mathura
[2] Department of Mechanical Engineering MNNIT, Allahabad

Abstract. This paper presents a selection procedure for fluid film journal bearing by incorporating fuzzy approach. In this paper a fuzzy based selection model is proposed which can be applied for selection of hydrostatic, hydrodynamic and hybrid journal bearing. Selection criteria is formulated for the choice of space requirement, cost of bearing and load carrying capacity of the corresponding journal bearing. This approach provides a third dimension to the existing method of selection of bearing, which is dynamic and may be further refined to address to every individual needs. Therefore a fuzzy logic approach is used for decision making.

Keywords: Fluid film Journal bearing, fuzzy logic, Expert system.

1 Introduction

Bearing are machine elements that permits relative motion of two parts in one or more directions with a minimum of friction while preventing motion in the direction of applied load. Fluid film bearings are classified according to the manner in which load is supported, viz., hydrodynamic, hydrostatic and hybrid. An important feature of a selection procedure for fluid film bearings is the strategy for selecting the bearing type and configuration, its fluid feeding control devices. These basic decisions are usually made or considered at early stage of the design process [5]. Many researchers have discussed the advantages of hydrostatic, hydrodynamic and hybrid bearings in their research work [9-14].

Fuzzy logic was first proposed by Zadeh L.A. [1] of the University of California at Berkeley in a paper. He has elaborated his ideas in 1973 through a research work that introduced the concept of "linguistic variables", which in this article equates to a variable defined as a fuzzy set. Other research followed, by the first industrial application, a cement kiln built in Denmark.

Fuzzy has enabled researchers to quantify data which is generic in nature. Until now generic information cannot be measured. How will you decide how worthwhile it will be visiting a food joint? Good food , great place, nice offer are just not sufficient to help you in taking a decision, unless you have a scale to measure how good, how much nice etc. Unless information or data is precise we cannot work on it or utilize it. People interact among themselves through natural language which exists in numerous variations, but when they try to interact with machines and systems, they encounter vague and imprecise concepts which are easy to understand but difficult to interpret.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 311–320.
springerlink.com

For example the statement "Temperature today is 38^{o}C " does not explicitly state that today it's hot, and the statement "Today's temperature is 1.2 standard deviations about the mean temp for daytime in the month of May " is fraught with difficulties: would a temp 1.1999999 standard deviations above the mean be hot? [2, 6, 7]

In this paper, the development of a fuzzy based selection strategy is described for fluid film journal bearing. Fuzzy set theory thus aims at modeling imprecise, vague and fuzzy information. Computers cannot adequately handle such problems, because machine intelligence still employs sequential (Boolean) logic. The superiority of the human brain results from its capacity of handling fuzzy statements and decisions, by adding to logic parallel and simultaneous information sources and thinking processes, and by filtering and selecting only those that are useful and relevant to its purposes. The strategy includes the design of expert system, selection of membership function, input, output and a fuzzy rule base. The selection strategy has been implemented on the selection of fluid film bearing in this paper with one example also.

2 Expert Systems – An Overview

In ordinary Boolean algebra, an element is either contained or not contained in a given set. Fuzzy sets describe sets of elements or variables where limits are ill-defined or imprecise, the transition between membership and Non-membership is gradual, and an element can "more or less" belong to a set consider for instance the set of "costly bearing". In Boolean algebra, it is assumed that any individual bearing either belongs or does not belong to the set of costly bearing. This implies that the individuals will move from the category of "costly bearing" to the complementary set of "cheep bearing".

Fuzzy set theory allows for grades of membership. Depending on the specific application, one might for instance decide that bearing of cost more than Rupees 20 lakhs under definitely costly, while bearing having cost below Rs 1000 is definitely not costly, and that a bearing having cost Rs. 10,000 is "more or less" costly, or is costly with a grade membership of 0.3, on a scale from 0 to 1.

2.1 Methodology

The fuzzy logic expert system for selecting correct bearing, the following three input is taken for consideration i.e. index of cost of bearing, space required for bearing, load bearing capacity" concept. This index reflects the degree of vagueness in cost or elusiveness in the information furnished by the vendor and the information collected by the designer from various other sources. Fuzzy logic is a very powerful tool for dealing with human reasoning and decision making processes which involve ambiguity, approximation, inaccuracy, inexactness, inexact information, perception, qualitativeness, subjectivity, uncertainty, vagueness or sources of imprecision that are non-statistical in nature. By applying fuzzy logic, one can quantify the contribution of a set of information to various parameters in terms of fuzzy membership. During the past few decades, fuzzy logic has used as an attractive tool for various applications

ranging from household goods, finance, traffic control, automobile speed control, nuclear reactor, and earthquake detections etc.

2.2 System Architecture

Here develop a fuzzy logic based expert system for selection of correct fluid film bearing. Figure 1 shows the control mechanism of such system. The fuzzy logic based expert system consists of four components: fuzzifier, inference engine, defuzzifier, and the rule base. The role of fuzzifier is to convert a crisp input variable into linguistic variables. That is ready to be processed by the inference engine. The inference engine using the fuzzified inputs and the rules stored in the rule base process the incoming data and produces linguistic output. Once the output linguistic values are available, the defuzzifier produces the final crisp values from the output linguistic values.

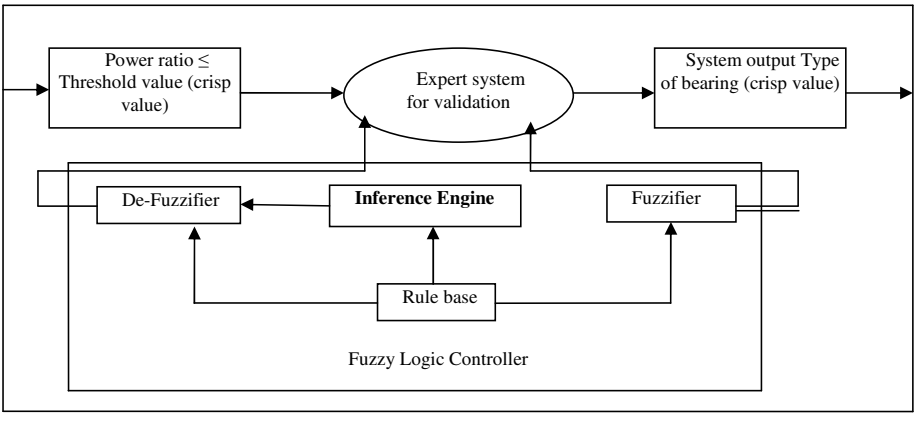


Fig. 1. Fuzzy expert System

The validation process starts by entering two sets of data; one furnished by the vendor and the other obtained by the designer reviewing the demand. This information is obtained from a standard design data catalogue form that is used by the designer. This form contains different sections, each section containing a set of information such as budget of company, available space of installation, load bearing requirement, type of machine in which bearing is installed e.g. pump , turbine etc. All the requirements are try to fulfill if the output values are below to a certain value. These qualitative measures are quantified and converted into linguistic variables with corresponding membership functions.

$$X_1 = \frac{\left(\sum_{i=1}^{I} \sum_{j=1}^{J} W_{ij} \Delta_{ij} \right)}{I} \tag{1}$$

Where W_{ij} is the weightage or impact factor given to the j[th] information of the i[th] section, and Δ_{ij} is a 0-1 variable ($\Delta_{ij} = 1$ if there is any deviation/difference in the information furnished by the vendor and the one obtained by the designer, 0 otherwise). It is worthwhile noting that the information that is crucial in decision making of selection of bearing is given higher weightage/impact factor. Also all the weights for a set of i[th] information, $\sum_{j=1}^{J} W_{ij}$ added to unity. Similarly, the values of the other inputs can be determined. The normalized values of these measures are used as inputs to the expert system. The degree of membership corresponding to a value of input is determined by the use of triangular membership functions because of their simplicity and good result obtained by simulation. These membership functions are designed on the basis of available information.

Figure 2 shows the definition of the fuzzy sets of the input and the output functions. A rule base is then constructed which will based on all the applicable input parameters and for each decision several rules are to be fired. Table 1 shows a sample rule base for the system under consideration which emphasize on the fact that in real life situations, the expertise of the human auditors will be used in the construction of the rule base. These rules result in an aggregate fuzzy set that represents a particular decision regarding the processing of the claims. This fuzzy set is then converted into a crisp number, which depicts the degree of suitability of the decision regarding the processing of the claims. The rules aggregation is done using weigthed average (WA) method. *Mamdani* implication is used to represent the meaning of "IF-THEN" rules. In this context, the statement "if X is A then Y is B" or A $\rightarrow$B results in a relation R such that $\mu_X(X,Y) = \min(\mu_A(X), \mu_B(Y))$. This implication is precise, computationally simple, and fits various practical applications. The *min* operator is a natural choice for the logical AND. Bellman and Giertz (1973) have devised a set of axioms that should be satisfied by the AND operator and have proved that *min* operator satisfies them.

The output of expert system is defuzzified on the basis of power ratio. Power ratio is the ratio of friction power and pumping power. It is denoted by K.

- $1 \leq K \leq 3$ Hydrostatic bearing.(Recess bearing)
- $3 \leq K \leq 12$ Hybrid bearing.(Non recessed bearing)
- $12 \leq K \leq 40$ Hydrodynamic bearing

Power ratio (K) = H_f / H_p

Where H_f = friction power ($\mu A_f U^2/h_0$), Hp = pumping power ($H_p = P_s.q$).

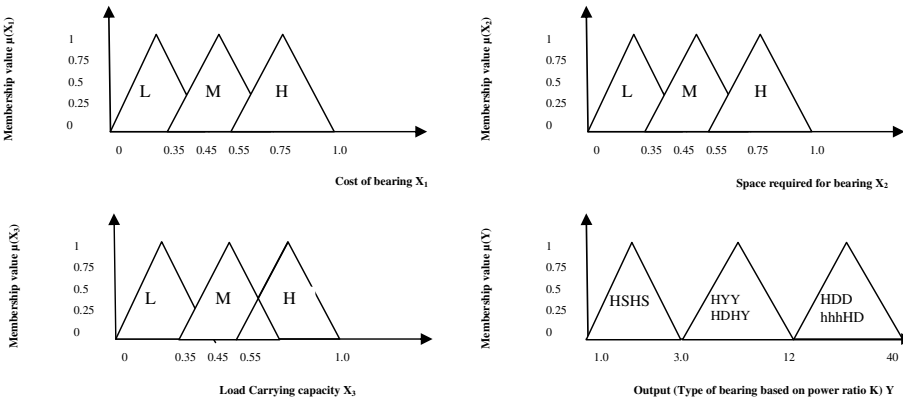


Fig. 2. Membership functions of inputs and outputs functions

Table 1. Sample rule base for the fuzzy logic based expert system

Rule No.	X_1 (I/P)	X_2(I/P)	X_3(I/P)	Y(O/P)	Rule No.	X_1(I/P)	X_2(I/P)	X_3(I/P)	Y(O/P)
1	Low	Low	Low	HD	15	Medium	Medium	High	HY
2	Low	Low	Medium	HD	16	Medium	High	Low	HD
3	Low	Low	High	HS	17	Medium	High	Medium	HY
4	Low	Medium	Low	HD	18	Medium	High	High	HY
5	Low	Medium	Medium	HS	19	High	Low	Low	HD
6	Low	Medium	High	HS	20	High	Low	Medium	HY
7	Low	High	Low	HD	21	High	Low	High	HY
8	Low	High	Medium	HY	22	High	Medium	Low	HD
9	Low	High	High	HY	23	High	Medium	Medium	HY
10	Medium	Low	Low	HD	24	High	Medium	High	HY
11	Medium	Low	Medium	HD	25	High	High	Low	HD
12	Medium	Low	High	HS	26	High	High	Medium	HY
13	Medium	Medium	Low	HD	27	High	High	High	HY
14	Medium	Medium	Medium	HS					

Where X_1 (I/P) cost of bearing, X_2 (I/P) Space required for bearing, X_3 (I/P) Load carrying capacity, Y (O/P) power ratio.

3 Algorithms – Using Fuzzy Approach

The steps of the expert system are summarized below:

1. **Input**. The crisp value of the cost of bearing, space required for bearing, load carrying capacity and other information obtained in the normalized form.

2. **Evaluate the main parameter**. Determine the cost of bearing index X_1, space requirement for bearing X_2, load bearing capacity X_3.
3. **Fuzzify the crisp values of inputs**. Through the use of membership functions defined for each fuzzy set for each linguistic variable (Figure 2), determine the degree of membership of a crisp value in each fuzzy set. Each of these three ambiguity indices have been divided into three fuzzy sets (LOW – L, MEDIUM – M and HIGH – H). The equations for computing memberships are:

$$\mu(X_I)_L = \begin{cases} \max\left\{0, \dfrac{X_I - a_I^L}{c_I^L - a_I^L}\right\} & if\ X_I < c_I^L \\ \max\left\{0, \dfrac{b_I^L - X_I}{b_I^L - c_I^L}\right\} & if\ c_I^L \le X_I \end{cases} \qquad \mu(X_I)_M = \begin{cases} \max\left\{0, \dfrac{X_I - a_I^M}{c_I^M - a_I^M}\right\} & if\ X_I < c_I^M \\ \max\left\{0, \dfrac{b_I^M - X_I}{b_I^M - c_I^M}\right\} & if\ c_I^M \le X_I \end{cases}$$

$$\mu(X_I)_H = \begin{cases} \max\left\{0, \dfrac{X_I - a_I^H}{c_I^H - a_I^H}\right\} & if\ X_I < c_I^H \\ \max\left\{0, \dfrac{b_I^H - X_I}{b_I^H - c_I^H}\right\} & if\ c_I^H \le X_I \end{cases} \tag{2}$$

where (a, c, b) are the vertices of the triangular membership function while L, M and H represents the fuzzy set LOW, MEDIUM, and HIGH, respectively.
4. **Fire the rule bases that correspond to these inputs**. All expert systems which is based on fuzzy logic uses IF-THEN rules. The "IF" part is known as antecedent or premise, whereas the "THEN" part is termed as a consequence or conclusion. Since all the three inputs have three fuzzy sets (LOW – L, MEDIUM – M and HIGH – H) therefore 27 (3x3x3) fuzzy decisions are to be fired. There are three outputs: Hydrostatic (HY), Hybrid (HD) and Hydrodynamic bearing (HD).
5. **Execute the inference engine**. Once all crisp input values have been fuzzified into their respective linguistic values, the inference engine will access the fuzzy rule base of the fuzzy expert system to derive linguistic values for the intermediate as well as the output linguistic variables. The two main steps in the inference process are aggregation and composition. Aggregation is the process of computing the values of the IF (antecedent) part of the rules while composition is the process of computing the values of the THEN (conclusion) part of the rules. During aggregation, each condition in the IF part of a rule is assigned a degree of truth based on the degree of membership of the corresponding linguistic term. From here, product (PROD) of the degrees of truth of the conditions are computed to clip the degree of truth from the IF part. This is assigned as the degree of truth of the THEN part. The next step in the inference process is to determine the degrees of truth for each linguistic term of the output linguistic variable. Usually, either the maximum (MAX) or sum (SUM) of the degrees of truth of the rules with the same linguistic terms in the THEN parts is computed to determine the degrees of truth of each linguistic term of the output linguistic variable.
6. **Defuzzification**. The last phase in the fuzzy expert system is the defuzzification of the linguistic values of the output linguistic variables into crisp values. The most

common techniques for defuzzification are center-of-maximum (CoM) and center-of-area (CoA). CoM first determines the most typical value for each linguistic term for an output linguistic variable, and then computes the crisp value as the best compromise for the typical values and respective degrees of membership. The other common method, CoA, or sometimes called center-of-gravity (CoG), first cuts the membership functions of each linguistic term at the degrees corresponding to the linguistic values. The superimposed areas under each cut membership function are balanced to give the compromised value. A disadvantage of this technique is the high computational demands in computing the areas under the membership functions. There are other variants of computing crisp values from linguistic values. These are mean-of-maximum (MoM), left-of-maximum (LoM) or smallest-of-maximum (SoM), right-of-maximum (RoM) or largest-of-maximum (LoM), weighted average (WA) and bisector-of-area (BoA) [7].

7. **Output of the decisions of the expert system**. In this case, the types of the outputs are: hydrostatic, hybrid bearing and hydrodynamic bearing. This selection is based on power ratio factor. The specific features of each controller depend on the model and performance measure. However, in principle, in all the fuzzy logic based expert system, we explore the implicit and explicit relationships within the system by mimicking human thinking and subsequently develop the optimal fuzzy control rules as well as knowledge base.

Example: For the purpose of illustration, Authors consider that a power generation company requires a fluid film bearing to support the turbine shaft and they provided three inputs as desired in fuzzy expert system. i.e. budget for bearing purchase is around Rs. 20 lakhs (cost of bearing)X_1, total area available for installation of bearing unit is approximately 70 m^2 X_2 and weight of turbine shaft is 500 kgf. X_3. These inputs represent the degree of vagueness/doubt in the information furnished during various time periods. The degree of vagueness/doubt in the information and the level of judgment used by the vendor as well as designer in deciding the type of bearings are always a challenge. At this type of situation fuzzy based expert system is a very good tool for decision making for both vendor as well as customers.

(1) First normalized all three inputs by dividing max value of corresponding input to the given input by the customer.
(2) Evaluate the authenticity. The values of the inputs have to be evaluated in fuzzy form, $X_1 = 0.40$; $X_2 = 0.70$ and $X_3 = 0.50$ (say).
(3) Fuzzification of the crisp values of inputs. Through the use of membership functions defined for each fuzzy set for each linguistic variable (Figure2), the degree of membership of a crisp value in each fuzzy set is determined as follows:

$$\mu(X_1)_L = \max\left\{0, \frac{b_1^L - X_1}{b_1^L - c_1^L}\right\} = 0.22 \qquad \mu(X_1)_M = \max\left\{0, \frac{X_1 - a_1^M}{c_1^M - a_1^M}\right\} = 0.25$$

$$\mu(X_1)_H = \max\left\{0, \frac{X_1 - a_1^H}{c_1^H - a_1^H}\right\} = 0 \qquad \mu(X_2)_L = \max\left\{0, \frac{b_2^L - X_2}{b_2^L - c_2^L}\right\} = 0$$

$$\mu(X_2)_M = \max\left\{0, \frac{b_2^M - X_2}{b_2^M - c_2^M}\right\} = 0.25 \qquad \mu(X_2)_H = \max\left\{0, \frac{X_2 - a_2^H}{c_2^H - a_2^H}\right\} = 0.667$$

$$\mu(X_3)_L = \max\left\{0, \frac{b_3^L - X_3}{b_3^L - c_3^L}\right\} = 0 \qquad \mu(X_3)_M = \max\left\{0, \frac{b_3^M - X_3}{b_3^M - c_3^M}\right\} = 1$$

$$\mu(X_3)_H = \max\left\{0, \frac{X_3 - a_3^H}{c_3^H - a_3^H}\right\} = 0$$

where

$(a_1^L, c_1^L, b_1^L) = (0, 0.225, 0.45)$; $(a_1^M, c_1^M, b_1^M) = (0.35, 0.55, 0.75)$; $(a_1^H, c_1^H, b_1^H) = (0.55, 0.775, 1.0)$

$(a_2^L, c_2^L, b_2^L) = (0, 0.225, 0.45)$; $(a_2^M, c_2^M, b_2^M) = (0.35, 0.55, 0.75)$; $(a_2^H, c_2^H, b_2^H) = (0.55, 0.775, 1.0)$

$(a_3^L, c_3^L, b_3^L) = (0, 0.225, 0.45)$; $(a_3^M, c_3^M, b_3^M) = (0.45, 0.50, 0.55)$; $(a_3^H, c_3^H, b_3^H) = (0.55, 0.775, 1.0)$

(5) Fire the rule bases that correspond to these inputs. Based on the value of the fuzzy membership function values for the example under consideration, the following rules apply:

Rule 5: If X_1 is LOW, X_2 is MEDIUM, X_3 is MEDIUM then Y is a Hydrostatic bearing (HS).
Rule 8: If X_1 is LOW, X2 is HIGH, X_3 is MEDIUM then Y is a Hybrid bearing (HY).
Rule 14: If X_1 is MEDIUM, X_2 is MEDIUM, X_3 is MEDIUM then Y is A Hydrostatic bearing (HS).
Rule 17: If X_1 is MEDIUM, X2 is HIGH, X_3 is MEDIUM then Y is a Hybrid bearing (HY).

(5) Execute the Inference Engine. We use the "root sum squares" (RSS) method to combine the effects of all applicable rules, scale the functions at their respective magnitudes. The respective output membership function strengths (range: 0-1) from the possible rules (R1-R27) are:

"Hydrostatic bearing index" = $\sqrt{\sum_{i \in AS} (\mu_{R_i})^2}$

$= \sqrt{(0.22)^2 + (0.25)^2} = 0.33$

"Hybrid bearing index" = $\sqrt{\sum_{i \in SF} (\mu_{R_i})^2}$

$= \sqrt{(0.22)^2 + (0.25)^2} = 0.33$

(6) Defuzzification. In this paper "fuzzy centroid algorithm" is used for defuzzification. The defuzzification of the data into crisp output is accomplished by combining the results of the inference process and then computing the "fuzzy centroid" of the area. The weighted strengths of each output member function are multiplied by their respective output membership function center points and summed.

Finally, this area is divided by the sum of the weighted member function strengths and the result is taken as the crisp output.
(7) Output of the decisions of the expert system. From Figure 3, it is concluded that the bearing should be hybrid bearing because the power ratio of the bearing is 6.5.

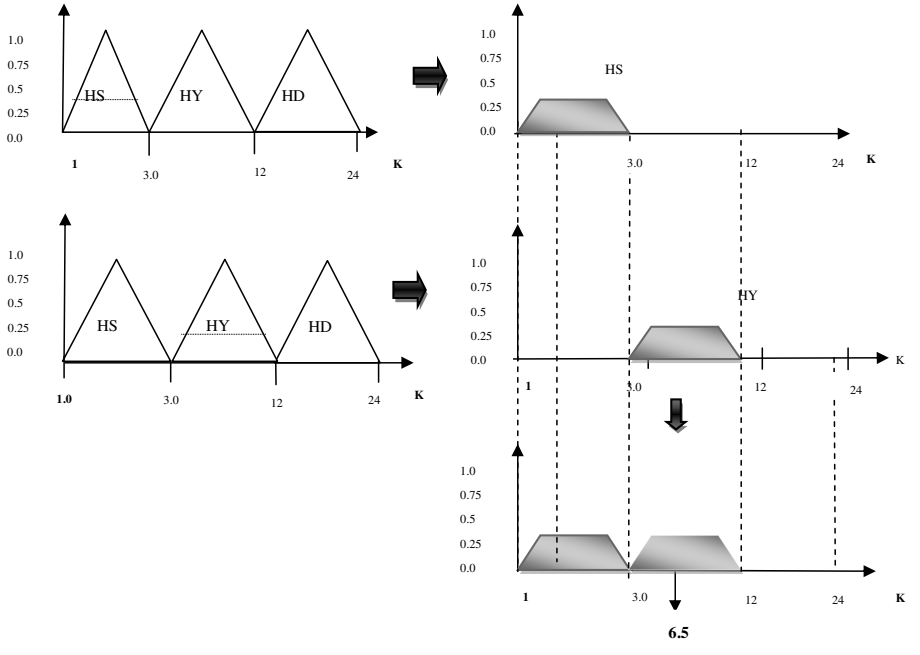


Fig. 3. Output of the decision of the expert system

4 Conclusions

The use of a fuzzy based expert system for selection of bearing is first time taken into consideration by the authors. Our future efforts will be on the improvement of the performance of the system by adjusting the membership function of the inputs. It would be interesting to tune the rule base using data from real life problems so that the performance of the system is optimized. We propose to use neural networks that can produce an optimum surface representing all the combination points from a few of the tested combinations.. It is worthwhile noting that inclusion of these factors would increase the size of the rule base to the point that the tuning of the rule base using data from real life scenarios will be deemed necessary to optimize the performance of the system. The system proposed through this work is evaluated on hypothetical data. This algorithm and methodology is also compatible with the continuous auditing paradigm.

References

[1] Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)
[2] Bellman, R., Giertz, M.: On the analytic formalism of the theory of fuzzy sets. Information Sciences 5, 149–156 (1973)
[3] Rowe, W.B.: Hydrostatic and Hybrid bearing Design. Butterworth Heinemann, Boston (1983)
[4] Row, W.B., Cheng, K., Ives, D.: An Intelligent Design System for Recessed Hydrostatic Journal Bearings. Wear 159, 95–105 (1992)
[5] Cheng, K., Rowe, W.B.: A selection strategy for the design of externally pressurized journal bearing. Tribology International 28(7), 465–474 (1995)
[6] Derrig, R., Ostaszewski, K.: Fuzzy techniques of pattern recognition in risk and claim classification. The Journal of Risk and Insurance 62(3), 447–482 (1995)
[7] Ross, T.J.: Fuzzy Logic with Engineering Applications. McGraw-Hill, Singapore (1997)
[8] Deshmukh, A., Lakshminarayana, T.: A rule-based fuzzy reasoning system for assessing the risk of management fraud. International Journal of Intelligent Systems in Accounting, Finance & Management 4, 231–241 (1998)
[9] Beek, A., Ostayen, R.A.J.: The design of partially grooved externally pressurized bearings. Tribology International 39, 833–838 (2006)
[10] Garg, H.C., Sharda, H.B., Kumar, V.: On the design and Development of Hybrid Journal Bearing: A Review. Tribotest 12, 1–19 (2006)
[11] Roy, L., Laha, S.K.: steady state and dynamic characteristics of axial grooved journal bearings. Tribology International 42, 754–761 (2009)
[12] Garg, H.C., Kumar, V., Sharda, H.B.: Performance of slot –entry hybrid journal bearings considering combined influences of thermal effects and non- Newtonian behavior of lubricant. Tribology International 43, 1528–1531 (2010)
[13] Sharma, S.C., Phalle, V.M., Jain, S.C.: Performance analysis of a multirecess capillary compensated conical hydrostatic journal bearing. Tribology International 44, 617–628 (2011)
[14] Phalle, V.M., Sharma, S.C., Jain, S.C.: Influence of wear on the performance of 2-lobe multirecess hybrid bearing system compensated with membrane restrictor. Tribology International 44, 380–395 (2011)

Implementation of New Biorthogonal IOFDM

A.V. Meenakshi, R. Kayalvizhi, and S. Asha

Periyar Maniammai University, Vallam, Thanjaur
meenu_gow@yahoo.com, Kayal2007@gmail.com,
ashasugumar@gmail.com

Abstract. Proposed biorthogonal interleaved OFDM system is used in Multiuser and multicarrier technique that has been recognized as an excellent method for high speed bi directional wireless mobile communication. In conventional interleaved OFDM system, convolution encoder is used as the channel encoder, but it leads to Bandwidth inefficiency and also reduces the throughput of the transmission and reception. The proposed bi orthogonal interleaved OFDM system is having the baud rate of 9600 kbps. This system is ultimately designed for the Bandwidth optimization and also it supports the Multi user transmission and reception of interleaved OFDM system.

Keywords: Biorthogonal modulation, Gold sequence generator, QPSK, QAM, IOFDM.

1 Introduction

In recent years, orthogonal frequency division multiplexing (OFDM) has been adopted as a standard for various applications like digital audio/video broadcasting (DAB/DVB), wireless LANs, etc. Conventional OFDM systems transform information symbol blocks and then insert redundancy in the form of either cyclic prefix (CP) or zero padding (ZP). The length of CP/ZP should be longer than the channel delay spread to avoid interblock interference (IBI) arising due to the frequency-selective nature of the channel. ZP assures symbol recovery even when channel nulls occur on some subcarriers, which is not possible with the use of CP. However, there is an increase in receiver complexity. The redundancy due to the CP/ZP causes reduction in the code rate of the communication system. IOFDM enhances the code rate without bandwidth expansion and without increasing the number of subcarriers but with a moderate increase in computational complexity and delay.

This paper is organized as follows. Section 2 presents the general block diagram of IOFDM system model with channel coding. In section 3 the above mentioned system model is explained with biorthogonal constant amplitude modulation instead of using convolutional encoder. Herewith describe the general block diagram of biorthogonal encoder with constant amplitude modulator. This explains the functional blocks of multi code generator, orthogonal parity vector matrix, constant amplitude encoder and orthogonal multiplier. This allows the system to adapt to the optimum data rate

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 321–331.
springerlink.com

vs. error rate for the current conditions. Section 4. Reviews the so called biorthogonal decoder with Walsh hadamard orthogonal code generator presented here. Section 5 discusses the simulation results of biorthogonal IOFDM system performance SNR Vs BER using matlab tool. Finally Section 6 describes the comparison results of simulations. Section 6 concludes the paper.

2 System Model

The block diagram in Fig. 1 describes the discrete-time baseband model of an IOFDM system [7]. It consists of channel encoder, IOFDM transmitter, AWGN channel, IOFDM receiver and channel decoder. The bit stream is mapped to an information symbol sequence using a modulation scheme like quaternary phase shift keying (QPSK). A transmitted block of length is formed as follows:

$$x(n) = [x(nN), ..., x(nN + N - 1)] \tag{1}$$

The ZP will avoid IBI between the transmitted blocks. After parallel-to-serial conversion of the sequence is then serially transmitted through a transmitting antenna.

At the receiver, the received sequence y(n) in the presence of noise is given by

$$y(n) = x(n) * h(n) + w(n) \tag{2}$$

Where w(n) denotes complex additive white Gaussian noise (AWGN). Here it is assumed that the impulse response is constant over the transmission of channel.

Fig. 2.1.

2.1 Bi-orthogonal Encoder

Convolutional codes are used extensively in numerous applications in order to achieve reliable data transfer including digital videos, radios, mobile and satellite communications. But in convolutional encoder, for 4 bits user data it generates 8 bits output which leads to Bandwidth in-efficiency. So we proposed Bi-orthogonal Encoder to overcome this problem. The block diagram for Bi-orthogonal Encoder is given by the figure 2.2.

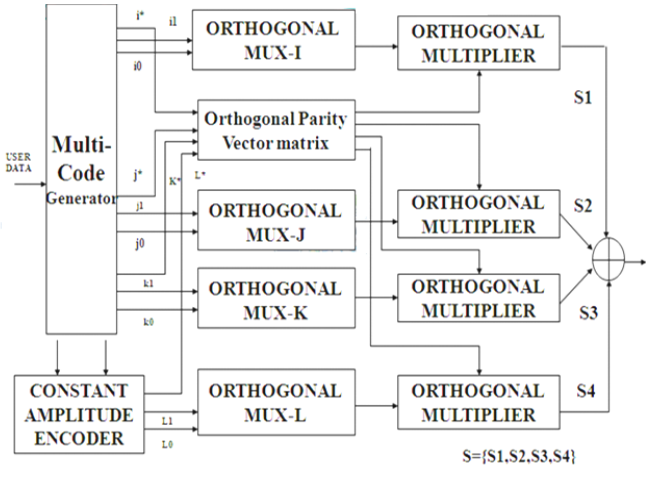


Fig. 2.2.

The proposed Bi-orthogonal Encoder gets 7 bits user data and generates 4 bits output. Thus it provides Bandwidth efficiency when compared with other encoders and also it reduces Bit Error Rate (BER). Multicode Generator consists of 2 blocks, Serial to Parallel Converter and Gold Sequence Generator. Serial to parallel Converter converts the data bits into number of branches according to the length of Gold Sequence. Gold Sequence Generator generates Gold Sequence by XOR the two Pseudo Noise sequences. The last two bits k1, k0 are always kept zero.

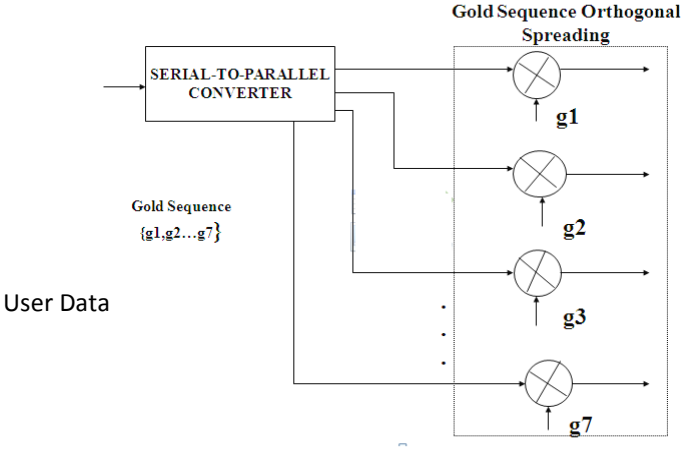


Fig. 2.3. Multicode Generator

The Gold Sequence Generator in Multicode Generator is given by,

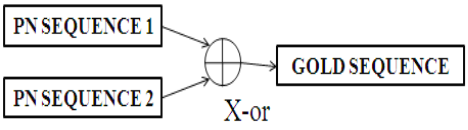


Fig. 2.4. Gold Sequence Generator

Three parity bits (L*,L1,L0) are generated from three groups of parallel bits (i*,i1,i0),(j*,j1,j0),(k*,k1,k0) by the constant amplitude encoder. According to the following equation

$$L^* = (I^* \wedge j^* \wedge k^*)^{(-1)} \tag{3}$$

$$L1 = i1 \wedge j1 \wedge k1 \tag{4}$$

$$L0 = i0 \wedge j0 \wedge k0 \tag{5}$$

Parity bits (L*, L1, L0) gets combined with the user data to reduce Bit Error Rate (BER). The Orthogonal Multiplier multiplies the outputs of Orthogonal MUX and Orthogonal parity vector matrix as shown in below, Where 'b' represents Orthogonal Parity Vector Matrix and 'C' is the Walsh-Hadamard Matrix.

$$\begin{aligned} S_i = b \begin{bmatrix} c_i \\ c_j \\ c_k \\ c_l \end{bmatrix} &= [i_* \;\; j_* \;\; k_* \; l_*] \begin{bmatrix} c_i \\ c_j \\ c_k \\ c_l \end{bmatrix} \\ &= i_* \cdot c_i + j_* \cdot c_j + k_* \cdot c_k + l_* \cdot c_l. \end{aligned} \tag{6}$$

Where i = 1,2,3,4.

The Bi-orthogonal Encoder produces the output S= {S1, S2, S3, S4}. This sequence is applied as input to IOFDM transmitter. The Transmitter block consists of aerial to parallel converter, IFFT, interleaver, parallel to serial converter. QPSK modulator, cyclic prefix adder**.** The resultant IOFDM signal is transmitted with additive white Gaussian noise. The corrupted IOFDM symbol is applied with IOFDM receiver. The IOFDM receiver consists of p/s converter, FFT, S/P convertor, deinterleaver and demodulator with cyclic prefix removal. The resultant symbol is fed with biorthogonal decoder.

2.2 Bi-orthogonal Decoder

The block diagram for Bi-orthogonal Decoder is given by,

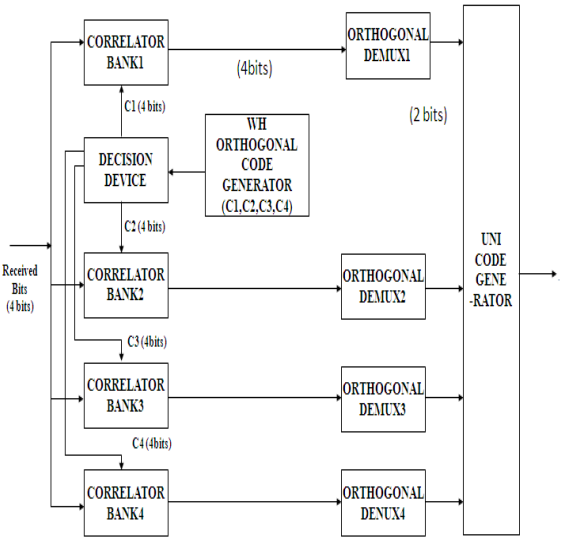


Fig. 2.5. Biorthogonal Decoder

A decoder is a device which does the reverse of an encoder, undoing the encoding so that the original information can be retrieved. The same method used to encode is usually just reversed in order to decode. It performs the reverse operation of Bi-orthogonal Encoder.

The procedure for generating Walsh Hadamard Orthogonal Code is shown in figure2.6.

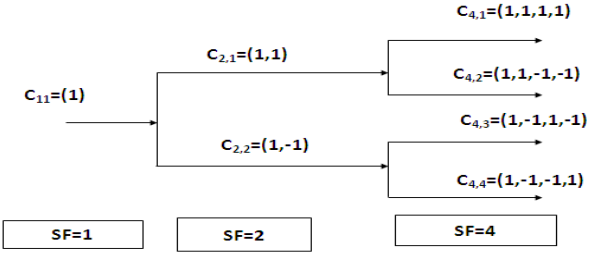


Fig. 2.6. Walsh Hadamard Orthogonal Code Generator

The generated Walsh Hadamard Code Matrix is given by,

$$H = \begin{vmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{vmatrix}$$

Unicode generator performs the reverse operation of Multicode generator. Its block diagram is given by,

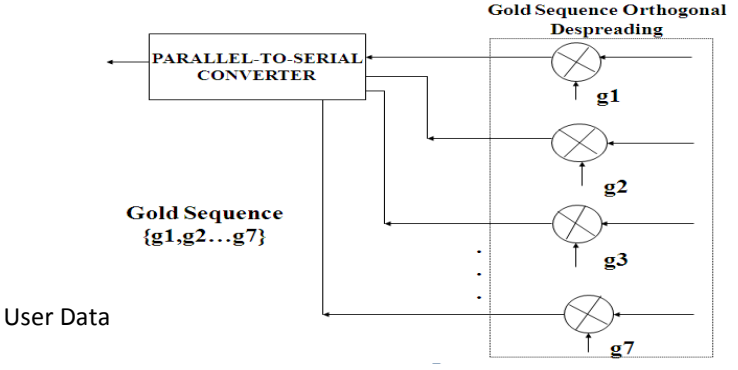


Fig. 2.7. Unicode Generator

The Unicode generator consists of 2-blocks, Parallel-to-Serial Converter and gold sequence despreader. The parallel to Converter converts the data bits in to number of branches according to the length of Gold Sequence.

3 Results

SPECIFICATIONS:

Bandwidth	100 MHz
Number of subcarriers	1024
OFDM/ symbol duration	12.5 μs (10.24 + 2.26: effective symbol +guard interval)
Packet length	0.6 ms (48 × 12.5 μs): 48 OFDM symbols
encoding	Convolution channel encoding
Modulation	QPSK

Channel model: AWGN and Rayleigh fading channel

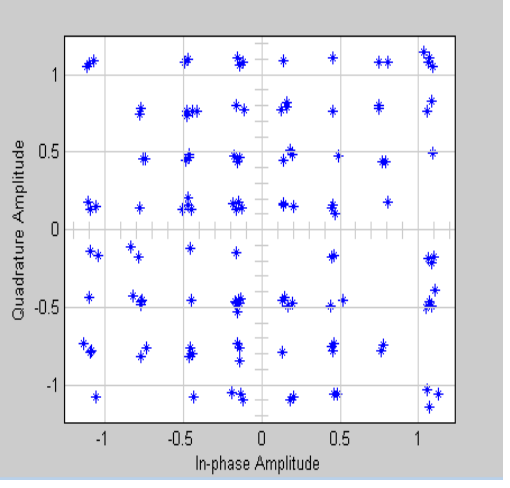


Fig. 3.1. Received signal constellation

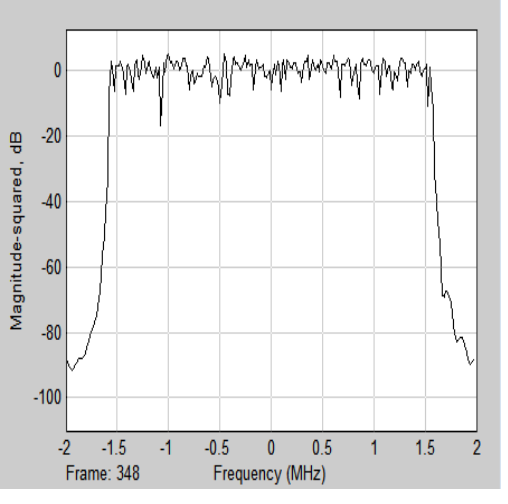


Fig. 3.2. Spectrum of Received signal

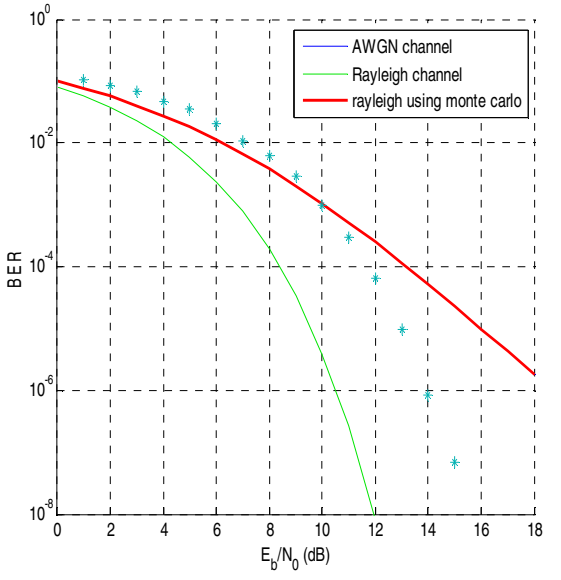


Fig. 3.3. SNR Vs BER performance

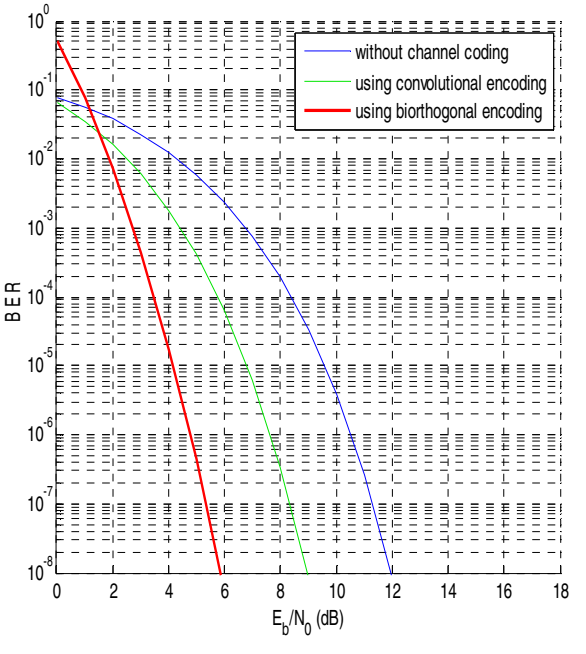


Fig. 3.4. SNR Vs BER performance of proposed system

Table 3.1.

S.N	Parameters	Convolution Encoder	Biorthogonal encoder
1	Bandwidth utilization	Minimum	Optimum Half when compared to convolution encoder
2	Data Rate	Upto 1MHz	Upto 1MHz
3	SNR	0-25 dB	0-25dB
4	Power	High (5v DC)	Low 3.3v DC
5	BER	$5.57E^{-138}$	$1.61E^{-60}$
6	Design Complexity	$5.57E^{-138}$	High

Table 3.2.

S.NO	**E_b/N_o(db)**	**BER**	**No. of bits used**
1.	0	0.123	900
2.	5	0.0356	3000
3.	10	$9.77E^{-4}$	102300
4.	15	$6.99E^{-8}$	$1.000002E^{8}$
5.	20	0.0	$1.000002E^{8}$
6.	**25**	**0.0**	**$1.000002E^{8}$**

Table 3.3.

Specification	E_b/N_0(db)	BER
Without channel coding	0.18	$1.31E^{-29}$
With convolution coding	0.18	$1.61E^{-60}$
With Biorthogonal	0.18	$5.57E^{-138}$

Table 3. shows the comparison between the convolution and Bi orthogonal encoder in IOFDM system. In this Bi-orthogonal Encoder, the BER from 0.5dB to 0.28dB for the SNR value of 0 – 25dB,date rate is improved from 128kb/s to 256kb/s and also the power consumption is reduced from 5v to 3.3v dc in bi_orthogonal encoder when compared to Convoltional Encoder. Bandwidth utilization is maximum as compared to convolutional encoder i.e its reduced to half. As design point of view the bi_oprthogonal encoder is somewhat complex than convolutional encoder.

4 Conclusion

The proposed implementation of Bi-orthogonal IOFDM System performs well in Wireless Environment and also recognized as an excellent method for high speed bi-directional wireless mobile communication. Here, we utilize the bandwidth efficiently and simultaneously the BER is also reduced, which has a great advantage when compared to the previous systems. And this system is ultimately designed for the Bandwidth Optimization and also it supports Multi-User Transmission and Reception of OFDM System. Simulation results verify that the performance of the IOFDM system in terms of BER is very close to that of the conventional OFDM system.

References

1. Damen, M.O., Chkeif, A., Belfiore, J.C.: Lattice code decoder for space-time codes. IEEE Commun. Lett. 4, 161–163 (2000)
2. Wang, J.-T.: Signal Detection for the STFC-OFDM System over Time Selective Fading Channels. IEEE, Zhang, Y., Wang, J., Song, J., Yang, Z.-X.
3. Wang, J.-T., Yang, Z.-X., et al.: Design of space-time-frequency transmitter diversity scheme for TDS-OFDM system. IEEE Trans. Consumer Electronics 51(3), 759–764 (2005)
4. Lee, K.F., Williams, D.B.: A space-time coded transmitter diversity technique for frequency selective fading channels. In: Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop, Cambridge, MA, pp. 149–152 (March 2000)
5. Stuber, G., Barry, J., Mclaughlin, S.W., Li, Y., Ingram, M.A., Pratt, T.G.: Broadband MIMO-OFDM wireless communications. Proc. of the IEEE 92, 271–294 (2004)
6. Tarokh, V., Jafarkhani, H., Calderbank, A.R.: Space-time block coding for wireless communications: performance results. IEEE J. Select Areas in Communications 17(3), 451–460 (1999)
7. Viterbo, E., Boutros, J.: A universal lattice code decoder for fading channels. IEEE Trans. Inform. Theory 45, 1639–1642 (1999)

8. Gong, Y., Letaief, K.B.: Space-Frequency-Time Coded OFDM for Broadband Wireless Communications. IEEE Trans. Center for Wireless Information Technology Dept. of Electrical & Electronic Engineering Hong Kong Univ. of Science & Tech.
9. Zheng, F.C., Burr, A.G.: Signal detection for orthogonal spacetime block coding over time-selective fading channels: a PIC approach for the i g systems. IEEE Trans. Commun. 53(6), 969–972 (2005)
10. Keller, T., Hanzo, L.: Adaptive Multicarrier Modulation: A Convenient Framework for Time-Frequency Processing in Wireless Communications. IEEE Proceedings of the IEEE 88, 609–640 (2000)
11. Wang, Z., Giannakis, G.B.: Wireless Multicarrier Communications. IEEE Signal Processing Magazine, 29–48 (May 2000)
12. Bingham, J.A.C.: Multicarrier Modulation for Data Transmission: An Idea Whose Time Has Come. IEEE Communications Magazine, 5–14 (May 1990)
13. Naguib, A.F., Seshadri, N., Calderbank, A.R.: Increasing Data Rate over Wireless Channels. IEEE Signal Processing Magazine, 76–92 (May 2000)

Application of Real Valued Neuro Genetic Algorithm in Detection of Components Present in Manhole Gas Mixture

Varun Kumar Ojha[1], Paramarta Dutta[1], Hiranmay Saha[2], and Sugato Ghosh[2]

[1] Department of Computer & System Sciences
[2] Visva Bharati, Santiniketan, West Bengal, India
[3] Centre of Excellence for Green Energy & Sensors System
[4] Bengal Engineering & Science University, West Bengal, India

Abstract. The article deals with the implementation of an Intelligent System for detection of components present in manhole gas mixture. The detection of manhole gas is important because the manhole gas mixture contain many poisonous gases namely Hydrogen Sulfide (H_2S), Ammonia (NH_3), Methane (CH_4), Carbon Dioxide (CO_2), Nitrogen Oxide (NO_x), and Carbon Monoxide (CO). A short exposure to any of these components with human beings endangers their lives. A gas sensor array is used for recognition of multiple gases simultaneously. At an instance the manhole gas mixture may contain many hazardous gas components. So it is wise to use specific gas sensor for each gas component in the gas sensor array. Use of multiple gas sensors and presence of multiple gases together result a cross-sensitivity. We implement a real valued neuro genetic algorithm to unravel the multiple gas detection issue.

Keywords: Cross-Sensitivity, Gas Sensor Array, Real Value, Neuro Genetic Algorithm

1 Introduction

In this article our focus is on implementing a real valued neurogenetic algorithm for development of an intelligent sensory system for detection of proportion of components present in manhole gas mixture. The manholes are built across the sewer pipeline. The sewer pipeline network is built in urban areas for draining out waste products. For cleaning and maintaining sewer pipelines several manholes built across this and persons have to get into these manholes to serve this purpose. In recent days, few deaths including municipality labourers and pedestrian are reported due to poisonous manhole gas exposure. This situation enforces us to mould our research involvement in this direction. In order to provide an intelligent sensory system, a neural network based system has to be developed such that it can act like an intelligent agent who can report the presence of poisonous gas component into the manholes. In [11,16,19] authors offers methods for manhole gas detection. In this article the training of the neural network is done by the real valued genetic algorithm. Where real valued genetic algorithm

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 333–340.
springerlink.com

searches out the best possible combination of synaptic weights for the neural network. A system like this will help labourers to being alert about the poisonous gases before entering into the manholes. The manhole gas mixture mainly contains Hydrogen Sulfide, Ammonia, Methane, Carbon Dioxide, Nitrogen Oxide, etc [1,2]. A sensor array containing distinct semiconductor based sensors report the presence of gases according to their concentration in manhole gas mixture. Reported values by the sensor array are incorporating cross-sensitivity which will be filtered out during the training of neural network.

2 Mechanism

2.1 The Gas Detection System Overview

The most general model of intelligent system for manhole gas mixture is shown in the Fig 1. In [17,18] the concerned authors presents their view on gas detection model. It is evident that gas mixture sample is collected in a gas mixture chamber and gas components in that mixture are in unknown proportion. The mixture is then allowed to pass over the gas sensor array and response of each sensor element is observed. The sensor array outcome is then fed to the data pre-processing block. And then pre-processed data is fed to the neuro genetic block for neural network training process. An alarm/report generator operation is based on the output of the neural network.

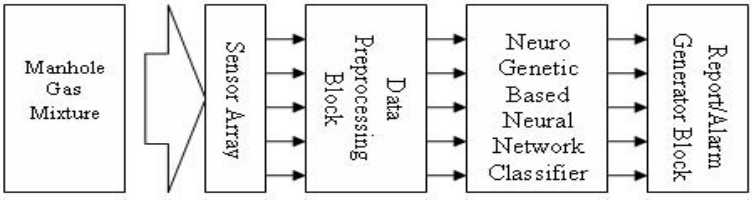


Fig. 1. Mixed Gas Detection System

2.2 Data Collection and Preprocessing

In the data collection process, we collect the sample of data for known gas mixture. The known gas mixture is a mixture of gas in the known concentration. To prepare a data sample we present the known gas mixture to gas sensor array and the sensor responses were taken. In this way we prepare several samples. A typical example of this data sample is shown in the Table 1. If we can focus on the first and second samples of Table 1,due to cross-sensitivity each sensor responses were increased in spite of increased in the concentration of single gas.

In the pre-processing block we normalize data samples according to equations 1 and 2. According to equation 1 the normalized value of gas H_2S in the sample

Table 1. Data Sample for gases in mixture on taking the known conc.

	Gas Mixture of Known Conc.					Sensor Response				
Sample	NH_3S	CO	H_4S	NO_2	CH_4	NH_3S	CO	H_4S	NO_2	CH_4
1	50	100	100	100	**2000**	0.053	0.096	0.065	0.037	0.121
2	50	100	100	100	**5000**	0.081	0.108	0.074	0.044	0.263
3	50	100	100	200	2000	0.096	0.119	0.092	0.067	0.125
4	50	100	200	200	5000	0.121	0.130	0.129	0.079	0.274
5	50	100	200	400	2000	0.145	0.153	0.139	0.086	0.123

2 is given by 100/5000 where the 100 appearing in the numerator is the conc. of the H_2S gas itself and the 5000 in the denominator is the max concentration of among all the samples. Similarly the sensor response are also normalized according to equation 2. Thus the input vector in neural network training pattern is the normalized sensor response and the target vector of the training pattern is the normalized value of concentration of gases in the gas mixture.

$$NC_{si} = \frac{C_{si}}{C_{max}} \tag{1}$$

Where NC_{si} is the normalized concentration of gas i of sample s, C_{si} is the actual conc. of gas i of sample s and C_{max} is the max value among all gases among all samples.

$$NR_{si} = \frac{R_{si}}{R_{max}} \tag{2}$$

Where NR_{si} is the normalized response of gas i of sample 2, R_{si} is the actual response of gas i of sample 2 and R_{max} is the max value of response among all gases across all samples.

The system output is denormalized to report the systems output in terms of concentration of gas components present in manhole gas mixture.

3 Our Neuro Genetic Approach

In the neuro genetic approach the neural network is trained using genetic algorithm [6,7,9]. The Genetic Algorithm is search algorithm based on the mechanics of natural selection and natural genetics [4,5]. The genetic algorithm searches out the optimal combination of synaptic weights for the neural network. In this approach we are using two layered feed forward neural network [3]. Where the input layer is containing 5 nodes, hidden layer is containing 3 nodes (reason why we choosing 3 nodes is discussed later) and the output layer contains 5 nodes. The 5 nodes each in input and output layer indicates the number of gas sensor in the sensor array. Moreover it means that our system will detect only 5 gases in the gas mixture.

3.1 Real Valued Neuro Genetic Algorithm

We are using the real valued genetic algorithm for training of the neural network. So coupling of neural network with genetic algorithm forms a concept of neuro genetic algorithm [10,12]. The flow chart of this neuro genetic algorithm is provided in the Fig 2.

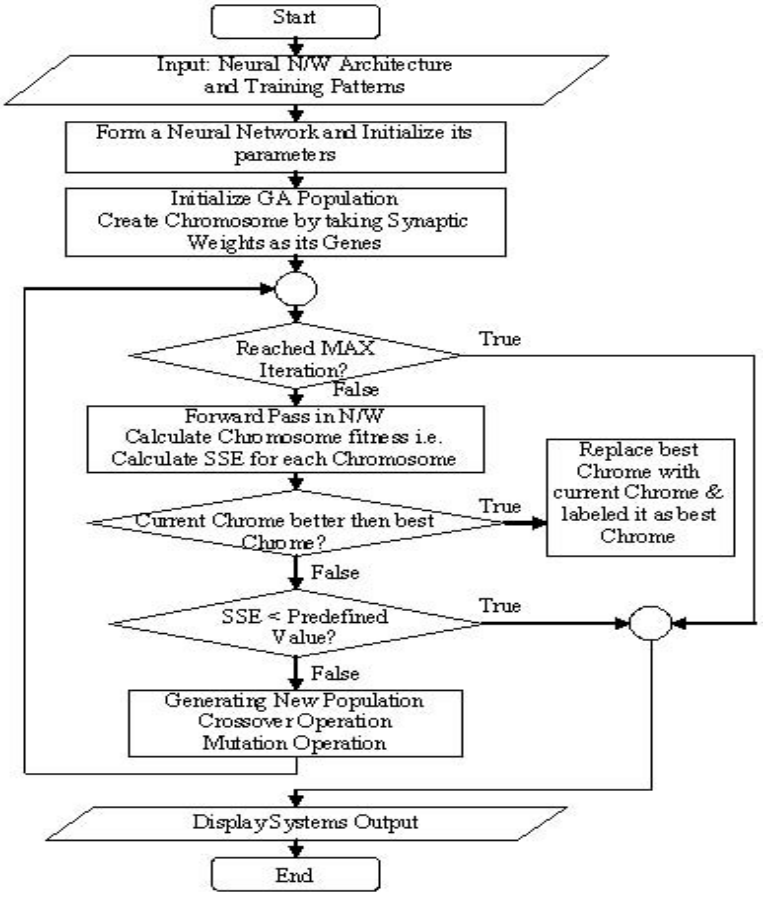


Fig. 2. Flowchart of the real valued neuro genetic algorithm

This Real Valued Neuro Genetic Algorithm offers minimization problem where it tries to minimize the sum squared error induced by the neural network by searching optimal synaptic weights of the neural network. The real valued genetic algorithm operates on the real value (float value). The chromosome in the real valued genetic algorithm is created using float values. In our approach we have chosen 32 bit IEEE 754 floating point format for representing a float value [13,14].

A single gene in the chromosome is a 32 bit IEEE 754 floating point format. Thus taking many gene together forms chromosome structure. The chromosome structure is discussed later. Subsequently, crossover and mutation operations are discussed.

Chromosome Structure. The Chromosomes used in the Real Valued Neuro Genetic Algorithm are created using the synaptic weights of the neural network [8]. Each synaptic weight (a float value) is considered as a gene in the chromosome and each gene is represented (encoded) as 32 bit IEEE 754 floating point format. Thus chromosome length in bits is multiple of 32. Let a neural network is having total of N synaptic weights ($W_1, W_2, W_3, W_4, \cdots, W_{N-1}, W_N$). Then the chromosome has N number of genes and length of the chromosome in number of bits is 32×N. The encoding of synaptic weights into chromosome structure is shown in the Fig 3.

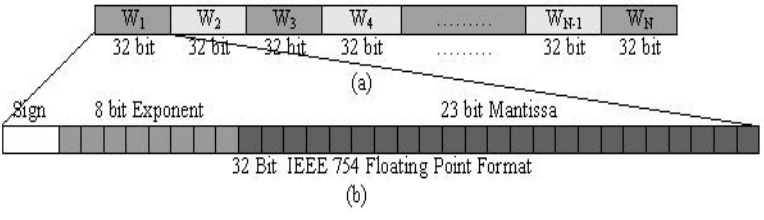


Fig. 3. *(a)* Encoding Synaptic Weights into Chromosome *(b)* The Sketch of Single Gene (W_i)

Where the synaptic weights $W_1, W_2, W_3, W_4, \cdots, W_{N-1}, W_N$ are the float values and its corresponding 32 bit binary representation is taken as a gene into the chromosome. And these genes will again decode into float value form the 32 bit IEEE 754 floating point formats according to equation 3.[13,14]

$$FloatValue(f) = (-1)^S \times Base^{E-127} \times 1.M \tag{3}$$

Selection Operation. For chromosome selection, the fitness proportionate selection (FPS) has been used.

Composite Single Point Crossover Operation. The crossover operation used in this case is a composite single point crossover vide Fig 4. The composite single point crossover is different from multipoint crossover.[1] Unlike multipoint crossover in the composite single point crossover, we uniformly fragment the

[1] In multipoint crossover the crossover operation is performed between the genes i.e. the exchange of genes happen between two chromosome at multiple points.

chromosome into N parts and the crossover performed only within that fragment at a randomly chosen index point. Here, we choose each gene as a separate fragment. The crossover operation is performed between similar positioned fragments (genes) of two chromosomes around a randomly chosen index point. A gene is 32 bit IEEE floating point format so crossover in each fragments is performed at a random index within 32 bit. We set crossover probability to 0.8.

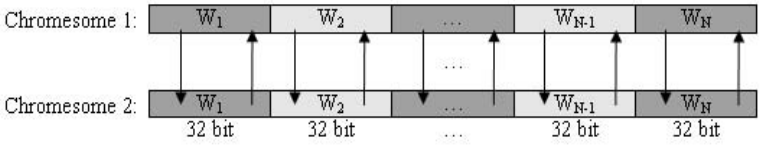


Fig. 4. Composite Single Point Crossover

While performing crossover operation we face a NAN (not a number) error problem. The value NaN is used to represent a value that does not represent a real number. NaN's are represented by a bit pattern with an exponent of all 1's and a non-zero fraction [13,14].

To avoid the NAN problem we choose a restrictive crossover vide Fig 5. In this technique we restrict the crossover to be done only within the last 29 or 30 bits i.e. the first 3 or 2 bits are not taking part in the crossover operation.

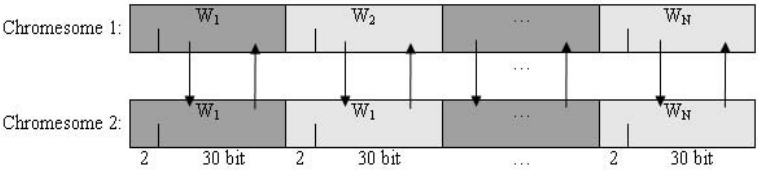


Fig. 5. Restrictive Composite Single Point Crossover

Composite Single Point Mutation Operation. The composite single point mutation is an operation which flips a randomly chosen bit into each fragment of the chromosome. Here in this case mutation in performed with low probability p = 0.2. Similar to the crossover operation the mutation is also performed within only last 29 bits or 30 bits to avoid NAN problem.

Fitness Function. As it is a neuro genetic approach, fitness function is the error function of the neural network. So sum squared error (SSE) of the network act as fitness functions for the chromosome. The chromosomes decoded as synaptic weights. Upon decoded synaptic weights SSE is computed using the equation 4.

$$SSE = \frac{1}{2} \sum_{p} \sum_{i} (O_{pi} - t_{pi})^2 \tag{4}$$

Where, O_{pi} and t_{pi} are the actual and desired outputs respectively retained at the output layer, p is the input pattern vector and i is the number of nodes in the output layer[15].

Stopping Criteria. The algorithm terminates on either of these two conditions.

1. The value of SSE reaches to an acceptable minimum.
2. The iteration number reached to maximum declared iteration.

3.2 Performance Analysis

The real valued neuro genetic algorithm is implemented in programming language JAVA and executed in the JDK 1.6 environment. The performance of algorithm are shown using iteration v/s SSE graph vide Fig 6(a). The neural network architecture is based on the performance of algorithm vide Fig 6(b).

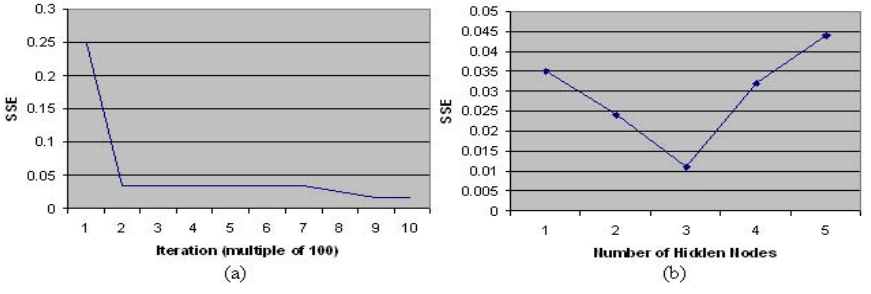


Fig. 6. *(a)* Performance of Network *SSE* against *iterations (b)* Performance of Network *SSE* against *Number of nodes in hidden layer*

4 Conclusion

In this article we have shown development of an intelligent sensory system composed of semiconductor based sensor array and real valued neuro genetic algorithm to provide solution to manhole gas detection issue. In the present article we provide a detail study of neuro genetic algorithm.The cross-sensitivity is treated during the preparation of data sample and training of neural network classifier.

Acknowledgments. Authors wish to acknowledge Department of Science & Technology, Govt. of India for the financial support vide Project No.: IDP/INP/02 to carry out this research.

References

1. Barsky, J.B.: Simultaneous Multi-Instrumental Monitoring of Vapors in Sewer Headspaces by Several Direct-Reading Instruments. Environmental Research 39(2), 307–320 (1986)
2. Hutter, G.M.: Reference Data Sheet on Gas(es), Meridian Engineering & Technology (November 1993), http://www.meridianeng.com/sewergas.html
3. Haykin, S.: Neural Network a Comprehensive Foundation, 2nd edn. Pearson Prentice Hall (2005)
4. Goldberg, D.E.: Genetic Algorithms in search, Optimization & Machine learning, 1st edn. Pearson Education (2006) ISBN 81-7758-829-X
5. Mitchell, M.: An Introduction to Genetic Algorithms. First MIT Press paperback edition (1998) ISBN 0262631857
6. Sindhu, S.S.S., Geetha, S., Sivanath, S.S.: A Neuro-genetic ensemble Short Term Forecasting Framework for Anomaly Intrusion Prediction. IEEE (2006) 1-4244-0716-8/06/$20.00
7. Kwon, Y.-K., Moon, B.-R.: A Hybrid Neurogenetic Approach for Stock Forecasting. IEEE Transactions on Neural Network 18(3) (May 20)
8. Srivastava, A.K., Srivastava, S.K., Shukla, K.K. In: Search of A Good Neuro-Genetic Computational Paradigm. IEEE (2000) 0-7803-5812-0/00/$10.009
9. Srivastava, A.K., Srivastava, S.K., Shukla, K. K.: On The Design Issue of Intelligent Electronic Nose System. IEEE (2000) 0-7803-581 2-0/00/$10.00
10. Barrios, D., Carrascal, A., Manrique, D., Rios, J.: Cooperative binary-real coded genetic algorithms for generating and adapting artificial neural networks. Springer-Verlag London Limited (2003)
11. Ojha, V.K., Dutta, P., Saha, H.: Detection of proportion of different gas components present in manhole gas mixture using backpropagation neural network. In: International Conference on Information & Network Technology (in press, 2012)
12. Li, H.-Q., Li, L.: A novel hybrid real-valued genetic algorithm for optimization problems. In: International Conference on Computational Intelligence & Security (2007)
13. Stallings, W.: Computer Organization and Architecture, pp. 222-234. Macmillan Publishing Company, ISBN 0-02-415480-6
14. IEEE Computer Society, IEEE Standard for Binary Floating-Point Arithmetic, IEEE Std. 754-1985
15. Sivanadam, S.N., Deepa, S.N.: Principles of Soft Computing, 1st edn. Wiley India (p) Ltd. (2007) ISBN 10:81-265-1075-7
16. Ojha, V.K., Dutta, P., Saha, H., Ghosh, S.: Linear regression based statistical approach for detecting proportion of component gases in manhole gas mixture. In: International Symposium on Physics and Technology of Sensors (in press, 2012)
17. Wongchoosuk, C., Wisitsoraat, A., Tuantranont, A., Kerdcharoen, T., Wisitsoraatb, A.: Portable electronic nose based on carbon nanotube-SnO_2 gas sensors and its application for detection of methanol contamination in whiskeys. Sensors and Actuators B: Chemical,SNB-12243
18. Tsirigotis, G., Berry, L.: Neural Network Based Recognition, of CO and NH_3 Reducing Gases, Using a Metallic Oxide Gas Sensor Array. In: Scientific Proceedings of RTU. Series 7. Telecommunications and Electronics, vol. 3 (2003)
19. Ojha, V.K., Dutta, P., Saha, H., Ghosh, S.: A Neuro-Swarm Technique for the Detection of Proportion of Components in Manhole Gas Mixture. In: International Conference on Modeling, Optimization and Computing (in press, 2012)

Concept Adapting Real-Time Data Stream Mining for Health Care Applications

Dipti D. Patil, Jyoti G. Mudkanna, Dnyaneshwar Rokade, and Vijay M. Wadhai

MAEER's MIT College Of Engineering, Pune, India,
Assistant Professor, Comp. Engg. Dept.,
dipti.dpatil@yahoo.com

Abstract. Developments in sensors, miniaturization of low-power microelectronics, and wireless networks are becoming a significant opportunity for improving the quality of health care services. Vital signals like ECG, EEG, SpO2, BP etc. can be monitor through wireless sensor networks and analyzed with the help of data mining techniques. These real-time signals are continuous in nature and abruptly changing hence there is a need to apply an efficient and concept adapting real-time data stream mining techniques for taking intelligent health care decisions online. Because of the high speed and huge volume data set in data streams, the traditional classification technologies are no longer applicable. The most important criteria are to solve the real-time data streams mining problem with 'concept drift' efficiently. This paper presents the state-of-the art in this field with growing vitality and introduces the methods for detecting concept drift in data stream, then gives a significant summary of existing approaches to the problem of concept drift. The work is focused on applying these real time stream mining algorithms on vital signals of human body in health care environment.

Keywords: Real-time data stream mining, concept-drift, vital Signal processing, Health Care.

1 Introduction

Data streams flow in and out from a computer system *continuously* and with varying update rates. They are *temporally ordered, fast changing, massive, and potentially infinite*.[1][11] It may be impossible to store an entire data stream or to scan through it multiple times due to its tremendous volume. So there is a need of analyzing this continuous data online without the overhead of storing it on a disk.

There exists a dynamic and promising field called data stream mining and knowledge discovery. To acquire knowledge base from raw data, emphasis is placed on innovative data stream mining concepts and techniques. This paper contains the general architecture of real-time data stream mining systems (RT-DSMS), different types of concept adapting algorithms, and finally finding useful patterns or knowledge from real-time data. Data streams are with the characteristics dynamic, non stationary, continuous, large volume, unstoppable, infinite.

The advanced research domain in DSM system is to handle concept drift in real-time data. While processing the data noise, errors, unwanted data, missing values

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 341–351.
springerlink.com

have to be removed. There are many proposed classification algorithms for concept drifting data streams. These algorithms support multidimensional analysis and decision making. Additional data analysis techniques are required for in-depth analysis, characterization of data changes over time. In addition, huge volumes of data can be accumulated beyond databases and data warehouses. Fig 1 shows the general data stream mining process. In applications like video surveillance, weather forecasting, telecommunication, sensor networks, satellites, call records, vital signals monitoring; data stream mining plays a key role to analyze the continuous data. The effective and efficient analysis of this data in such different forms becomes a challenging task. Also the issue of memory constraints has to handle as enormous data is generated continuously.

Expert system technologies, which typically rely on users or domain experts *manually,* input knowledge into knowledge bases. Unfortunately, this procedure is prone to biases and errors, and is extremely time-consuming and costly. **R**eal **T**ime **D**ata **S**tream **M**ining (RT-DSM) process performs data analysis and may uncover important data patterns, contributing greatly to business strategies, knowledge bases, and scientific and medical research. Hence, advanced RT-DSM algorithms are discussed in this paper and how they can be applied on vital signals of human body for health care is depicted.

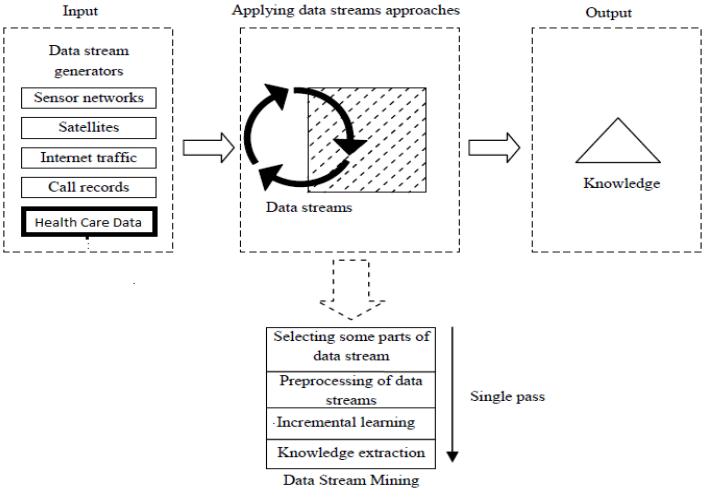


Fig. 1. General Process of Data Stream Mining

Section 2 discusses various issues in real-time data stream mining. Work related to Stream mining process and algorithms is described in section3. Section 4 explains algorithmic strategies for RT-DSM. The stress is on Ensemble-Based Classifiers. Section 5 depicts how these mining techniques can be applied on vital signals of human body for health care predictions and some of the results after applying data mining techniques on health care data in real-time. Section 6 represents Conclusion and future work.

2 Issues in Real Time Data Stream Mining

Issues and challenges beneath the concept drift are discussed below [1] [2]:

A. **Robustness issue:** The noise problem is more crucial for stream data mining, because it is difficult to distinguish noise from changes caused by concept drift. If an algorithm is too eager to adapt to concept changes, it may over fit noise and might be interpreting it as data from a new concept. If an algorithm is too old fashioned and slow to adapt, it may overlook important changes.

B. **Adaptation issue:** The concept generating a data stream drifts with time due to changes in the environment. These changes cause the model learned from old data is obsolete, and model updating is necessary.

C. **Performance issue:** To assure on-line responses with limited resources, continuous mining should be "fast and light", that is:
   a. Learning should be done very fast, preferably in one pass of the data;
   b. Algorithms should make light demands on memory resources.

D. **Sampling data from a stream:** Value or set of values at a point in time and/or space.

E. **Filtering a data stream:** Extract only the specific data that you want to see, and then display it in the manner that you want to see it. To address these issues, analysis of distinct algorithms and strategies is required with modest resource consumption.

The core assumption when dealing with the concept drift problem is uncertainty about the future. If it is assumed that the source St of the target instance St+1 is not known with certainty, it can be assumed, estimated or predicted but there is no certainty. Otherwise the data can be decomposed into two separate data sets and learned as individual models or in a combined manner.

3 Related Work

Background theory of real-time data stream mining process is explained below:

A] Concept drift: The underlying concept changes over time, so the learner should adapt to this change. It degrades the accuracy of classification system up to a point that the expected quality. Concept drift occurs during the classification mining process of data stream. Accuracy has been used to detect concept drift, which is sensitive to noise and affected by the effectiveness of the chosen classification algorithm.

B] Concept-evolution: New classes evolve in the stream, which makes classification difficult. This change can be real or virtual. Concept drifts can be grouped into two main families: *abrupt* and *gradual*. The abrupt type refers to situations where changes can be modeled suddenly and gradual drift is step-like changes affecting the environment in which the classification system is deployed. Gradual models situations where the process slowly evolves over time.

C] Change Detection Algorithms: Designing more efficient, accurate and parameter-free methods to detect change, maintain sets of examples and compute statistics. Trying to prove that the framework and the methods are useful, efficient and easy to use is tedious job. The imminent need for turning raw data into useful information and knowledge augments introduces development of systems, algorithms and frameworks that address streaming challenges. The storage, querying and mining of such data sets are highly computationally challenging tasks. Mining data streams is concerned with extracting knowledge structures represented in models and patterns in non stationary streams of information. Generally, two main challenges are designing fast mining methods for data streams and need to promptly detect changing concepts and data distribution because of highly dynamic nature of data streams.

D] Concept drift adaptation process: Concept drift refers to the learning problem where the target concept to be predicted changes over time in some unforeseen behaviors. It is commonly found in many dynamic environments, such as data streams, P2P systems, etc. Real-world examples include network intrusion detection, spam detection, fraud detection, epidemiological, and climate or demographic data, etc. Figure 3.1 depicts the incremental process of single learning instance [11], where the training model evolves with the concept drift and accordingly testing is carried out.

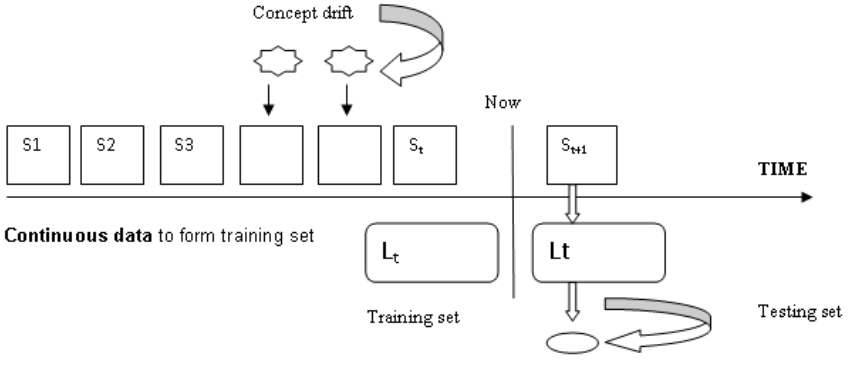


Fig. 3.1. Incremental Learning of Single Learning Instance

4 Algorithmic Strategies

Moving towards the innovative ideas, selecting the algorithm which is proper and efficient in giving results is very important.

Algorithms: General reasons for selecting the algorithms:[3]

- Popularity
- Flexibility
- Handling high dimensionality
- Applicability.

There are different categories to adopt concept drifts.

Following categories include different algorithms.

1. Clustering
2. Decision Trees
3. Ensemble based classification

We are focusing on ensemble based classification algorithm. A classifier is said to predict randomly, if the probability of data point x being classified to a class c is equal to c's class distribution in the current data block [4].

Algorithm 1: ***Ensemble Building***

Input: Data stream with class labels available intermittently,
= Permitted error

= Classifier Precision

Output: A set of classifiers, Global set G,
G= {C1, C2, C3,.., Cn }
Classification of Testing data
1. Global set← G { };
2. Ensemble set ← G { };
E ← { };
Training
3. New Classifier Required← true;
4. E ← { };
5. Get data chunk T from input stream with class label
6. MaxMSE ← classification error for data set T using a classifier predicting random
7. CEi ← classification error for data set T using
Classifier Ci
8. **for** classifier Ci in G
9. **if** CEi <

10. New Classifier Required ← false;
11. GO TO *Training*
12. **endif**
13. **endfor**
14. **for** classifier Ci in G
15. **if** CEi < (AcceptanceFactor * MaxMSE)
16. E ← {E} Union { Ci }
17. Wi ← MaxMSE - CEi
18. **endif**
19. **endfor**
20. CE Ensemble ← Calculate Classification error using ensemble set E with weights
21. **if** (CE Ensemble <)

22. New Classifier Required ← false;
23. GO TO *Training*
24. **endif**

25. **if** (New Classifier Required = false)
26. Ci ← Build Classifier (Data Chunk T, Classifier Precision α)
27. G ← {G} Union {Ci}
28. GO TO *Training*
29. **endif**

End *Training*

Testing

30. Classify the incoming stream using ensemble set E until the next set of la beled data is available.

Algorithm 2: ***Training the Dynamical Discriminative Model***

At the initial stage, the algorithm uses the *iterative reweighted least squares* method to train a logistic regression classifier based on an off-line data set. Then the parameter vector of the initial classifier is used as the initial value of *wt*. Based on the same data set, the initial value of covariance of the measurement noise and parameter vector can also be estimated online. Because the covariance of *wt* which denotes the degree of concept drift could change over time, we estimate its value based on a length-fixed buffer when it is filled with new examples and then update the related equations periodically. The performance of a classifier is quite stable when the size of the buffer is varied [5].

Input: *S*: a dataset from the incoming stream
C: a off-line dataset for evaluating the initial value of parameters
K: a size-fixed buffer for estimating *a*
Output: *wt*: a series of parameter vector of classifier for each time stamp learn
the initial parameter vector *wt* from *C* using the
IRLS method;
estimate the value of *a* from *C* using ;
while *S not empty* **do**
    get an instance *xt* from *S*;
    compute the prior estimate for *wt* and *Pt*
    compute the posterior estimate for *wt* and *Pt* using
    output the posterior estimate of *wt*;
if *the buffer K is full* **then**
estimate the value of *a* from *K* using ;
Pt= *aI*;
empty *K*;
end

Algorithm 3: ***Adaptive Ensemble Classifier***

The ensemble classifier are built and updated in an online manner. Once a new training instance arrives, it is used to update the ensemble classifier in an online bagging scheme. We store the new-comer instance in the predefined-size evaluation set. If the evaluation set is full, the least recent instance in the evaluation set will be removed. Whenever a chunk (the chunk size is smaller than the evaluation set size) of new instances arrive, we perform ensemble reconstruction and subset selection. The most recent chunk of training data is used to evaluate the ensemble for the current concept, and decide how much component classifiers should be dropped and which to

be dropped. New classifiers are constructed and added to the ensemble to keep the ensemble size constant. Besides ensemble reconstruction, we choose a sub-optimal subset of the component classifiers that have the best accuracy in the evaluation set to participate in the final decision [6].

E: ensemble classifier N: Chunk-size
I: new instance M:Training set size ES: evaluation set

```
For (n = 1….M)
{
UpdateEnsemble (E, In);
UpdateEvaluationSet (In,ES);
If (n%N==0)
{
Reconstruct(E,ES);
SubsetSelection(E,ES)
}}
```

Algorithm 4: Online Bagging Algorithm

Bagging usually works better than the individual component classifier. We can expect better accuracy than a single classifier by using the bagging ensemble classifier. Compared with boosting, bagging is less affected by noise in the training data. It implies that bagging may work better than boosting in real world application. Modified bagging algorithm is given here: [7]

Inputs: ensemble E, Ensemble Size S, training example T
On-line learning algorithm for the ensemble members
OnlineBaseLearningAlg.
1. for t=1 to S do
2. K←Poisson
3. while K>0 do
4. hm= OnlineBaseLearningAlg(hm,T)
5. K=K-1
6. end while
7. end for
Output: updated ensemble h.

Algorithm 5: ***DWCDS: Double Window Based Classification in real time data stream mining***

Due to limited size of sliding window the number of samples consider are less and may not take into account the concept drift. To overcome this problem Double Window Based Classification in real time data stream mining (DWCDS) is proposed. It is based on the changes of the original data distribution in the window to detect concept drifts. Correspondingly, the window sizes are adjusted dynamically to enhance the adaptability to concept drifts. Experiments show that DWCDS performs better on the concept drift detection, the ability of robustness to noise and the accuracy of classification. [10]

Input: Training set DSTR

Test set: DSTE, Attribute set A, Maximum height of trees h0, Basic classifier count N, Capacity of each basic classifier K, Sliding Window SW, Minimum threshold of window MinSW, Maximum Threshold of window MaxSW, Basic window w, Minimum no. of split instances nmin, Coefficient of drift warning t1, Coefficient of drift t2;

Output: Error rate of classification

Procedure: DWCDS { DSTR, DSTE, A, h0, N, K, SW, MinSW, MaxSW, w, nmin, t1, t2}

{

For (i=1; i<(basic classifier count-N); i++)

Generate k-random decision trees as a basic classifier CTi using the data in Sliding Window;

While (a new instance arrives)

{

If (Sliding Window ==full)

T: Detect concept drifts using the data distribution changes of current streaming data in the sliding window (SW)

and a basic window (w);

If (a concept drift occurs)

T: 1. Delete the classifier CTi with the worst performance on classification from CT;

2. Build a new basic classifier using the data that has a new concept in SW and put it into CT;

3. Adjust the size of the sliding window;}}

For (each test instance E DSTE)

Classify it in a voting mechanism using ensemble classifier {CTi}

Return the error rate of classification;

}

The algorithms discussed above are applied and tested on different datasets. For verifying the applicability and flexibility of these algorithms the accuracy evaluation of these algorithms are given in Table1.

Table 1. Comparative analysis of RT-DSM algorithms

Sr. No.	Name of Algorithm	Accuracy	Dataset used
1.	Ensemble Building	95%	Nursery data
2.	Training the Dynamical Discriminative Model	93.45%	Synthetic data
3.	Adaptive Ensemble Classifier	97.23%	SEA dataset
4.	Online Bagging Algorithm:	More than 96%	Hyper-plane Dataset
5.	DWCDS	93.5 to 98%	KDDCup99 Dataset

5 Implementation and Results

An objective of a health process is one where patients can stay healthy with the support of expert medical advice when they need it, at any location and any time. An associated aim would be the development of a system which places increased emphasis on preventative measures as a first point of contact with the patient. As the vital signals plays key role for predicting health status of a human, and these signals are continuous in nature(ECG, EEG, Heart Rate, SPO2 etc.), we propose to apply RT-DSM on these signals. Some of the basic data mining algorithms like K-means and Graph Theoretic algorithm are applied on the vital signals like ABPdias(diastolic blood pressure), ABPsys(sysstolic), Heart rate, SPO2.

Publicly available clinical data have been selected to validate the effectiveness of the proposed framework. Focusing on the intensive care scenario, the multi-parameter intelligent monitoring for intensive care (MIMIC) database [13] contains nearly 200 patient days of real-time signals. We used 64 records from the numeric section of the database, which provides measures sampled at 1 Hz. For each record, the gender, age, and disease of the patient are known. Same signals can be taken in real-time with the help of wireless body sensors and behavior of these signals can be learned on the fly to predict the patient's health status with the help of above discussed ensemble-based classifiers.

The sample results are shown in fig 5.1 for SPO2 signal and for ABP (dias) signal in figure 5.2 with K-means algorithm. The clustering algorithms applied are offline in nature, so they will be used for formation of historical rule base. We propose to apply above discussed real-time data stream mining algorithms on the vital signals to dynamically update the rule base and predict the health risk accurately.

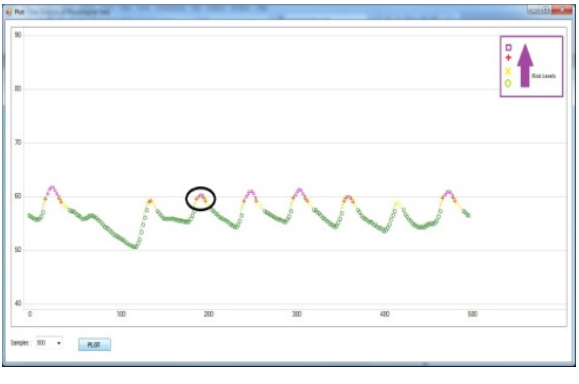


Fig. 5.1. SPO2 Signal to Analyze Risk Level

This will help to take immediate preventive actions in case high health risk. Also there will not be any need to keep patient in ICU in wired environment, instead patient can be continuously monitored from his home and aleretd in a risky situations.

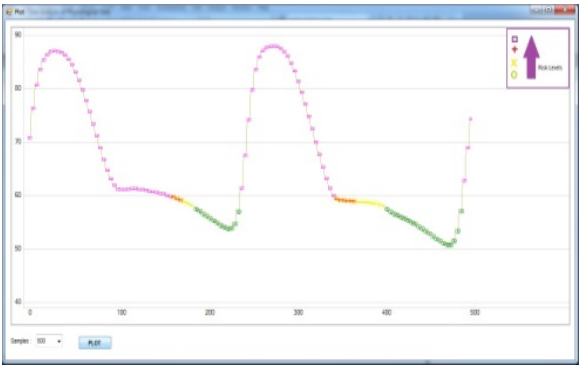


Fig. 5.2. ABP[dias] Signal to analyze Risk Level

6 Conclusion and Future Work

Different ensemble based classifier systems for RT-DSM has been discussed. All these methods are capable of performing any-time classification, learning in one scan and detecting drift in the underlying concept. The important issue of adapting concept drifts has been solved. We are focusing on real time data stream mining of health care data. The experiments are carried out on various vital signals of human body by applying the data mining algorithms like K-means and Graph Theoretic algorithm to predict the health risk level. The same data will be taken in real-time and dynamic algorithms will be applied on these vital signals to continuously monitor the health status.

References

[1] Harries, M.B., Sammut, C., Horn, K.: Extracting hidden context. Machine Learning 32(2), 101–126 (1998)

[2] Hulten, G., Spencer, L., Domingos, P.: Mining Time changing data streams. In: Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2001, vol. 1, pp. 97–106. ACM Press (2001)

[3] Ramamurthy, S., Bhatnagar, R.: Tracking Recurrent Concept Drift in Streaming data using Ensemble Classifiers. In: Proc. of Machine Learning and Applications, ICMLA 2007, pp. 404–409 (2007)

[4] Su, B., Shen, Y.-D., Xu, W.: Modeling Concept Drift from The Perspective of Classifiers. In: Proc. of Cybernetics and intelligent Systems, pp. 1055–1060. IEEE (2008)

[5] Wu, D., Liu, Y., Mao, Z., Ma, W., He, T.: An Adaptive Ensemble Classifier For Concept Drifting Stream. In: IEEE Symposium on Computational Intelligence and Data Mining, CIDM 2009, pp. 69–75 (April 2009)

[6] Minku, L.L., White, A.P., Yao, X.: The impact of diversity on learning in the presence of concept drift. IEEE Transactions on Knowledge and Data Engineering, 730–742 (2010)

[7] Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Jiawei Han University of Illinois at Urbana-Champaign (2006) ISBN 13: 978-1-55860-901-3 ISBN 10: 1-55860-901-6

[8] Bifet, A., Holmes, G., Kirkby, R., Pfahringer, B.: Data Stream Mining A Practical Approach (2011)
[9] Overview, Technical report, Vilnius University, 2009 techniques, related areas, applications Subjects: Artificial Intelligence arXiv:1010.4784v1 (cs.AI) Report number (October 2010)
[10] Kholghi, M.: An Analytical framework for data stream mining techniques based on challenges and requirements. In: IJEST (2011)
[11] The MIMIC database on Physio Bank (October 2007), http://www.physionet.org/physiobank/database/mimicdb
[12] Apiletti, D., Baralis, E., Bruno, G., Cerquitelli, T.: Real-Time Analysis of Physiological Data to Support Medical Applications. IEEE Transactions on Information Technology in Biomedicine 13(3), 313–321 (2009)

Efficient Domain Search for Fractal Image Compression Using Feature Extraction Technique

Amol G. Baviskar[1] and S.S. Pawale[2]

[1] Research Scholar, Vishwakarma Institute of Technology, Pune, India
[2] Asst. Professor, Vishwakarma Institute of Technology, Pune, India
amolbav@gmail.com,
sanjeshpawale@gmail.com

Abstract. Fractal image compression is a lossy compression technique developed in the early 1990s. It makes use of the local self-similarity property existing in an image and finds a contractive mapping affine transformation (fractal transform) T, such that the fixed point of T is close to the given image in a suitable metric. It has generated much interest due to its promise of high compression ratios with good decompression quality. The other advantage is its multi resolution property, i.e. an image can be decoded at higher or lower resolutions than the original without much degradation in quality. However, the encoding time is computationally intensive [8].

Image encoding based on fractal block-coding method relies on assumption that image redundancy can be efficiently exploited through block-self transformability. It has shown promise in producing high fidelity, resolution independent images. The low complexity of decoding process also suggested use in real time applications. The high encoding time, in combination with patents on technology have unfortunately discouraged results.

In this paper, We have proposed efficient domain search technique using feature extraction for the encoding of fractal image which reduces encoding-decoding time and proposed technique improves quality of compressed image.

Keywords: Range Blocks, Domain Blocks, Feature Vectors, Domain Search.

1 Introduction

Images are stored on computer as bits representing pixels, or points forming a picture element. Since the human eye can process large amounts of information (some 8 million bits), many pixels are required to store moderate quality images. These bits provide the "yes" and "no" answers to the 8 million questions that determine the image. For example, a single 800 by 600 pixel true-color image requires three bytes per pixel, plus a header, which amounts to over 1.37 Mb of disk space, thus almost filling a 1.4 Mb high-density diskette. Clearly, some form of compression is necessary. As well as saving storage space, compressed files take less time to transmit via modem, so money can be saved on both counts.

Fractal is first introduced in geometry field. The birth of fractal geometry is usually traced back to the IBM mathematician Benoit B. Mandelbrot and the 1977 publication

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 353–365.
springerlink.com

of his book, "The Fractal Geometry of Nature". Fractal image or video compression is a new compression method which is based on self-similarity within the different portions of the image. It was first promoted by M.Barnsley, who founded a company based on fractal image compression technology.[1,8]

2 Related Work

2.1 Image Compression

Image compression is the application of data compression on digital images. In effect, the objective is to reduce redundancy of the image data in order to be able to store or transmit data in an efficient form. Image compression can be lossy or lossless. Lossless compression is preferred for archival purposes and often medical imaging, technical drawings, clip art or comics. This is because lossy compression methods, especially when used at low bit rates, introduce compression artifacts. Lossy methods are especially suitable for natural images such as photos in applications where minor (sometimes imperceptible) loss of fidelity is acceptable to achieve a substantial reduction in bit rate.

Table 1. Image Compression Techniques[8]

Method	Advantages	Disadvantages
WAVELET	Higher Compression Ratio State-of-the-Art	Coefficient Quantization Bit Allocation
JPEG	Current Standard	Coefficient Quantization Bit Allocation
VQ	Simple Decoder No Coefficient Quantization	Slow Codebook Generation Slow Bpp
FRACTAL	Good Mathematical Encoding Frames Resolution-free Decoding	Slow Encoding Bit Allocation

2.2 Fractal Image Compression

The word *fractal* was derived from the Latin fractus, which means broken, or irregular and fragmented. Mandelbrot claimed that classical Euclidean geometry was inadequate at describing many natural objects, such as clouds, mountains, coastlines and trees.

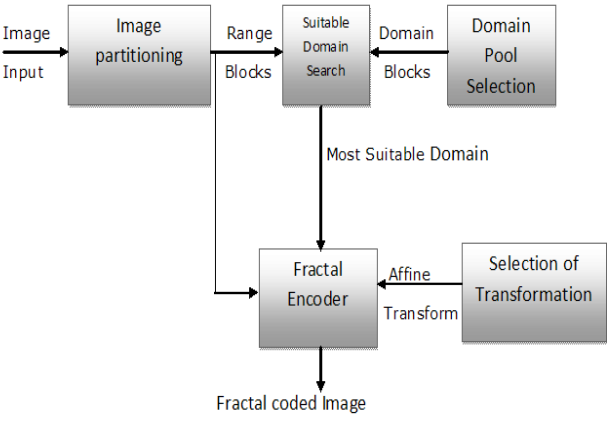


Fig. 1. Block diagram of Fractal Image Encoding Process

2.3 Geometric Transformation

We make use of eight isometrics that can be combined with the spatial contraction operator. They are: Four rotations 0^o, 90^o 180^o, 270^o and four flips over the vertical middle line, horizontal middle line, 45^o diagonal line and 135^o diagonal line.[8]

2.4 Fractal Image Encoding and Decoding Approach

The basic idea of fractal image compression is as following: divide initial image into small image blocks with non-overlapping Range blocks (R) and overlapping Domain blocks (D). For each R block, find domain block D which is the most similar to current R block by applying isometric transformation on D block. The blocks can be partitioned using any of 4 schemes:

1. Fixed-Sized Partitioning
2. Quad-Tree Partitioning
3. HV-Partitioning
4. Triangular Partitioning

Let us consider n*n square image (where n is width and height of the image) , r is the size of range block then the number of non overlapping range blocks will be (n*n/r*r) and number of overlapping domain blocks are $(n-2r+1)^2$.

As each range block is compared with each and every transformed domain block, which takes more amount of time for encoding. Hence it is not suitable for real time application. In the proposed work we have formulated technique to reduce number of domain blocks thereby reducing time for encoding and improving quality of compressed image.[17]

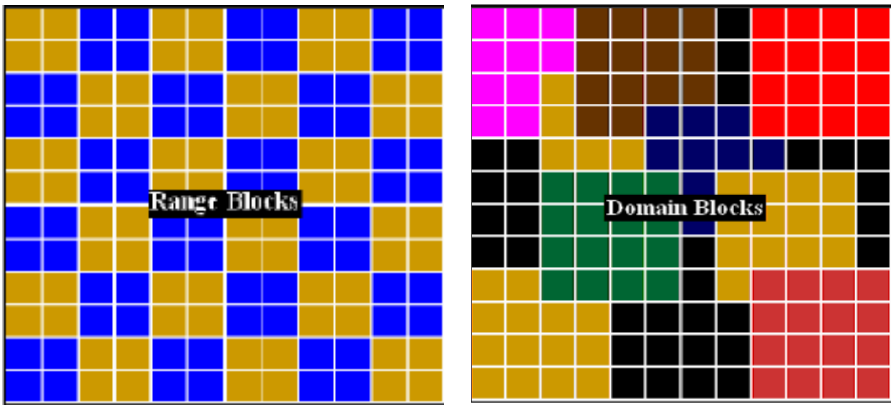


Fig. 2. Range and Domain block [5]

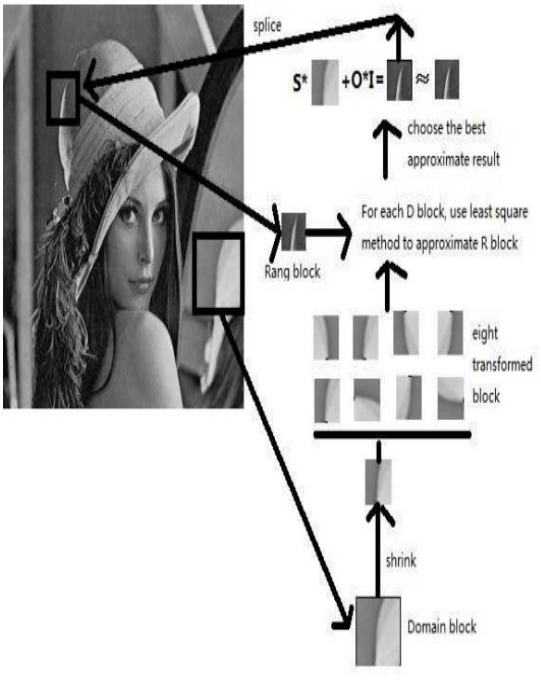


Fig. 3. The schematic diagram of the Fractal Encoding process

3 Proposed Technique

3.1 Reduction of Domain Block

Consider image of size 256 * 256. Size of Range block is 8 * 8. Hence, Size of domain block is 16 * 16.Total number of range blocks possible is: (256/8) * (256/8) = 32 * 32 = 1024.

These are overlapping blocks in image of size d=2 x r. and total number of domain blocks possible are [1,8]

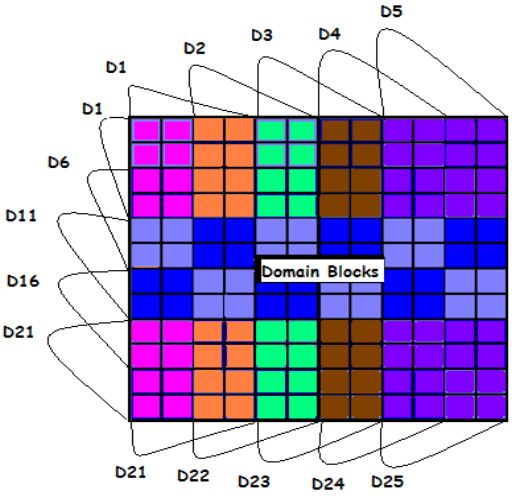


Fig. 4. Schematic view of reduction in domain block

Table 2. Domain Block Comparison[3,4,8]

For an Existing method : $(n-2r+1)^2$	**For an Proposed method :** $([n/r]-1)^2$
(256-2*8+1) * (256-2*8+1) = 58081	(256/8-1) * (256/8-1) = 961
Total number of affine transform for domain blocks are: 58081*8 = 464648	Total number of affine transform for domain blocks are: 961*8= 7688
Therefore, for a single range block 464648 matches must be done to select best domain block.	Therefore, for a single range block 7688 matches must be done to select best domain block.
So, for 1024 range blocks matches must be 1024 * 464648 = 475799552 are done.	So, for 1024 range blocks matches must be 1024*7688 = 7872512 are done.

For each R block, find a D block from D pool which is the most similar to it. The concrete steps are as following:

1) Shrink D block to the size of R block, marked D' block, and the specific shrinking method used is four neighborhood regional method.

2) Transpose, turned D' block. Specifically, we can choose eight affine transformations which proposed by Jacquin, and the corresponding transformation matrix. After transformation we get eight image blocks for each D' block.

3) After selection of suitable partitioning, domain-pool and transformation, third step of fractal encoding process is the search of suitable candidate from all available domain blocks to encode any particular range block. This step of fractal image compression is computationally very expensive, because it requires a large number of domain-range comparisons.

Comparison of range and domain block is very lengthy and time consuming process; in the proposed scheme we are using feature extraction method for the same.

3.2 Feature Extraction Technique

In suitable domain search we look for the compatibility between the range blocks and domain blocks. One way is to compare the images as a whole and another is to extract a few numbers of features that characterize the domain and range images. Then the comparison of range blocks and domain blocks is based on these features rather than on individual pixels. In this way the complexity of the problem is reduced, which results as fast coding process.

A few numbers of features (mean, standard deviation, skewness and kurtosis) of all the image blocks are calculated.[3,4,10]. Will give probability of histogram. Suppose we have a block of size 4 by 4,

```
101  105  98   89
103  110  103  96
104  110  110  93
107  118  110  98
```

we plot histogram of these Block as:

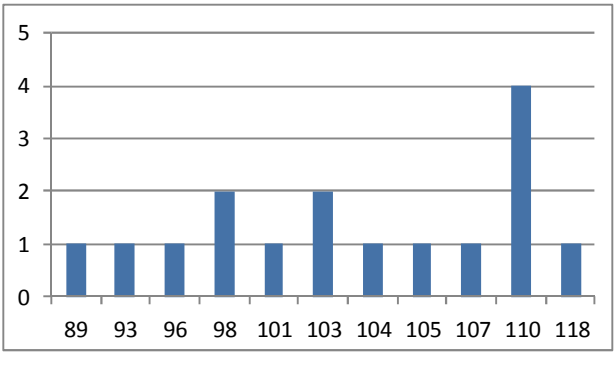


Fig. 5. Chart showing Histogram for Above Block

Here, y axis show number of pixels and on x axis shows k=0 to 255 intensity.
In this scheme whole process of suitable domain search is divided in three phases.

1) In first phase an average block (A) is computed, which is equal to the average of range blocks (r). Along with this a few number of features (mean, variance, skewness, kurtosis) of blocks are extracted and feature vector (f [m, v, s, k]) for

range, domain and average blocks are formed. Further operations of desired task are performed on these feature vectors.

2) In second phase, domain feature vectors (fd) are compared with average feature vector (fA), difference between these vectors is computed and stored in an array (Domain codebook).

3) In third phase, range feature vector (fr) corresponding to the range to be encoded is compared with average feature vector (fA) and difference between them is calculated.

Nearest value of this difference is searched in an array and domain block corresponding to nearest point is assigned to the particular range block. This technique will reduce the number of range-domain comparison with remarkable amount. Approximation is computed on the basis of vectors instead of images; hence the total time requirement of the desired process is reduced significantly.

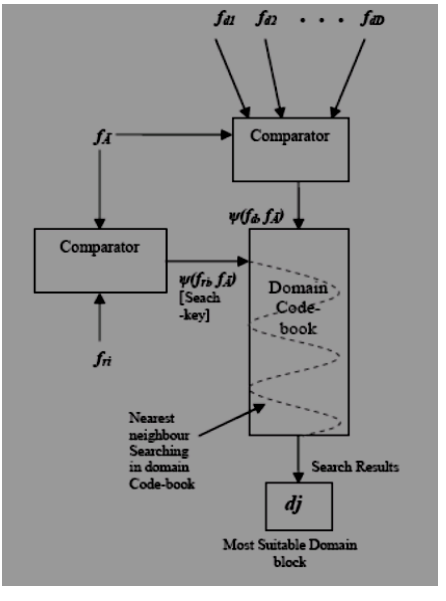


Fig. 6. Suitable Domain Search[4]

3.3 Proposed Algorithm

1) *Suppose image size is 12 * 12*
2) *Division of Range Blocks (e.g.:2 by2), therefore total number of range block = R1 to R36*
3) *Division of Domain Blocks (e.g.:4 by 4), therefore by our proposed technique we reduce total number of Domain Block that will be D1 to D25.*
4) *Compute the average image ($\bar{A}$), equal to the average of all range block.*

$$\bar{A} = \sum_{i=1}^{R} \frac{ri}{|R|}$$

Eg. For $\bar{A} = \sum_{i=1}^{36} \frac{ri}{36}$

5) *Calculate Histogram of Range Blocks , Domain Blocks and Average Image Block*

Extract Features :-

Mean (m) $= \sum_{k=1}^{K} k\, p_k$

Standard deviation (v) $= \sqrt{\sum_{k=1}^{K} (k-m)^2\, p_k}$

Skewness (s) $= \frac{1}{v^3} \sum_{k=1}^{K} (k-m)^3\, p_k$

Kurtosis (ku) $= \frac{1}{v^4} \sum_{k=1}^{K} (k-m)^4 \quad p_k - 3$

For all Range, Domain Blocks and Average Image Block

6) *Represent images as feature vectors (f),*
   Range feature vector $= f_r\,[m_r, v_r, s_r, ku_r]$
   Domain feature vector $= f_d\,[m_d, v_d, s_d, ku_d]$
   Average feature vector $= f_{\bar{A}}\,[m_{\bar{A}}, v_{\bar{A}}, s_{\bar{A}}, ku_{\bar{A}}]$
7) *Calculate Euclidian distance* $(\psi(f_d, f_{\bar{A}}))$ *for all domain blocks.*
8) *Store these errors in array*
9) *Calculate Euclidian distance* $(\psi(f_{ri}, f_{\bar{A}}))$.
10) *Search nearest value by comparing, while searching nearest value we calculate S, O. i.e brightness and contrast value of each domain block.*
    Here, comparing of
    $$\Psi(f_{ri}, f_{\bar{A}}) - \Psi(f_{di}, f_{\bar{A}}) = diff.$$
11) *Assign domain block corresponding to the nearest error value to the desired range block.*
12) *Get an encoded array*

3.4 Decoding an Image

The decoding is to find compressed image, by starting with any initial image. The procedure applies a compressed local affine transform on the domain block corresponding to the position of a range block until all of the decoded range blocks are obtained.

The procedure is repeated iteratively until it converges. The problems that occur in fractal encoding are the computational demands and the existence of best range-domain matches. The most attractive property is that image can be decoded at an enlarged size so that the compression ratio may increase exponentially [8]. However searching the domain pool is computationally intensive. For an n*n image, the number of range blocks are (n*n/r*r) and the number of domain blocks are $[(n/r)-1]^2$. The computation complexity for the best match between a range block and a domain block is O (r^2). If r is Constant, the computation complexity of entire search is O (n^4).

The reconstruction of an image from a fractal coding is a very simple process. One starts with any image of the same size as the original image. The transforms are then applied iteratively to their respective range blocks. After seven iterations, accuracy within .1 dB should be attained.

By adjusting the sizes of the blocks and giving the decoder a different size for the original image, we can reconstruct the image with the available amount of detail at

any resolution, without the additional pixilation introduced by various forms of interpolation. There will be a minor amount of blocking, but this becomes less important at higher resolutions.

4 Implementation

The proposed algorithm is implemented using JAVA. Coding simulations have been tested on 8-bit/pixel grayscale JPEG, PNG, BMP images. In pre-processing, the image to be encoded is partitioned in fixed size non-overlapping rectangular parts to form range blocks. This is done to increase the amount of self-symmetry. We have taken the range blocks of size 16 * 16, average block is also have same size as rang blocks. The domain blocks are formed by partitioning the image in overlapped blocks of double size as that of range blocks. The domain blocks of size 32 * 32 are used. We have evaluated the results on four test images, as shown in figure. Every image is taken in three sizes 512 * 512, 256 * 256 and 128 *128. Here we have computed the time consumed of suitable search in two conditions, first when the errors are stored in the form of an array and second when the errors are stored in a binary tree. The effect of size variation and tree incorporation on the performance is studied.

5 Evaluation and Results

All the Results are performed in AMD Turion64 with nVIDIA display. Various different images (50 to 60) are used to produce these benchmarks.

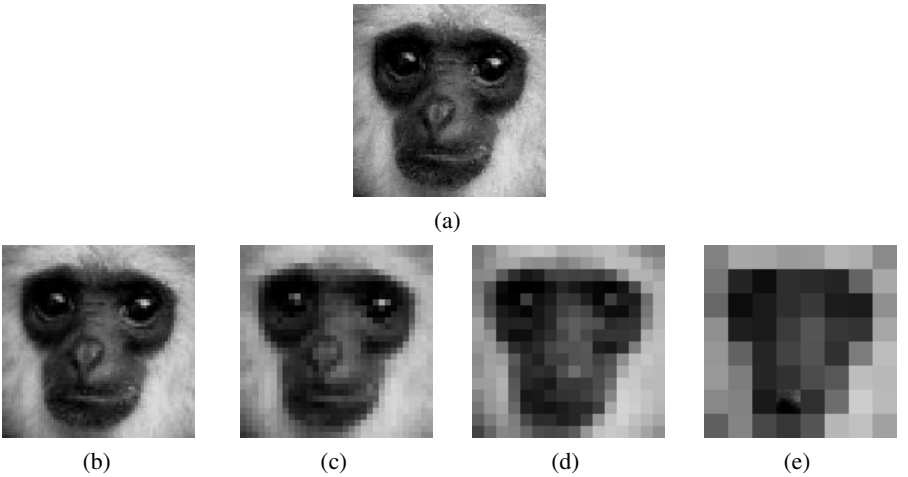


Fig. 7. (a) show original image for Monkey, decompressed images obtained for varying range block sized r=2,4,8,16 Figure 7(b), 7 (c), 7 (d), 7 (e)

Table 3. Shows various attributes related to image quality, encoding and decoding time for an for the particular images

Range	PSNR (DB)	CR	Encoding Time (ms)	Decoding Time (ms)
2	29.58143	5.7523017	112563	2390
4	24.890816	6.8705215	66500	422
8	21.075237	18.0	13656	235
16	17.743042	19.946447	3266	218

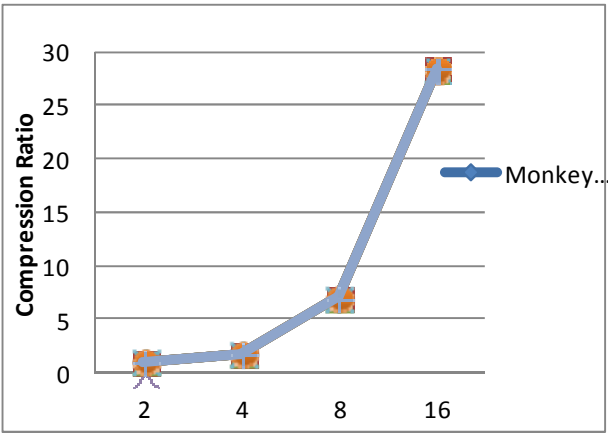


Fig. 8. Range block size Vs Compression ratio

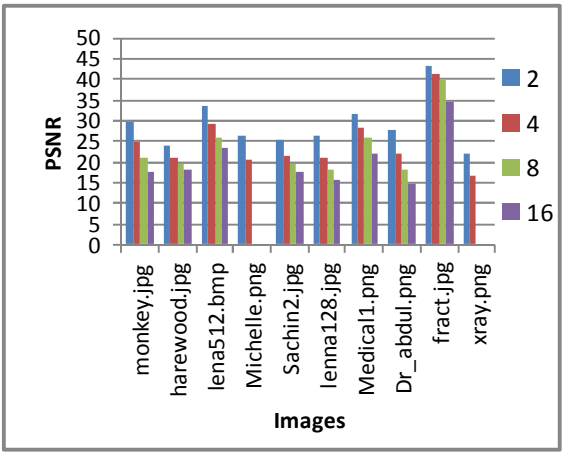


Fig. 9. PSNR Vs Images

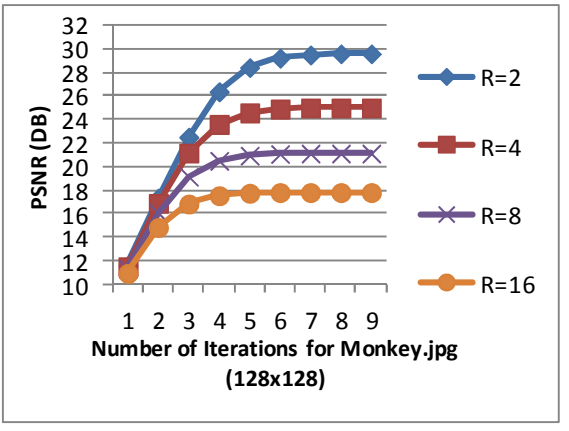


Fig. 10. Number of Iteration Vs PSNR

6 Proposed Flowchart

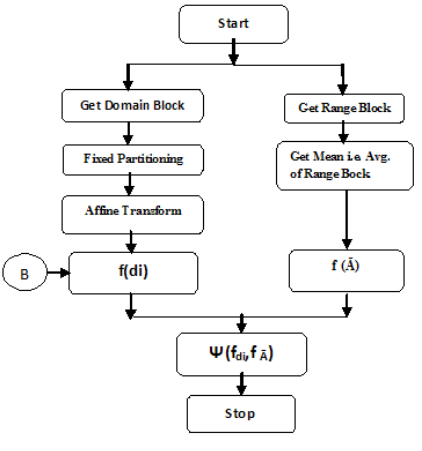


Fig. 11. Operation on Domain blocks

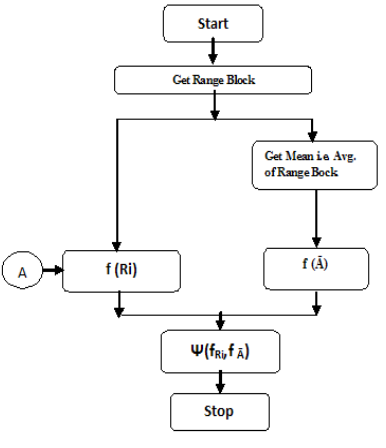


Fig. 12. Operation on Range Block

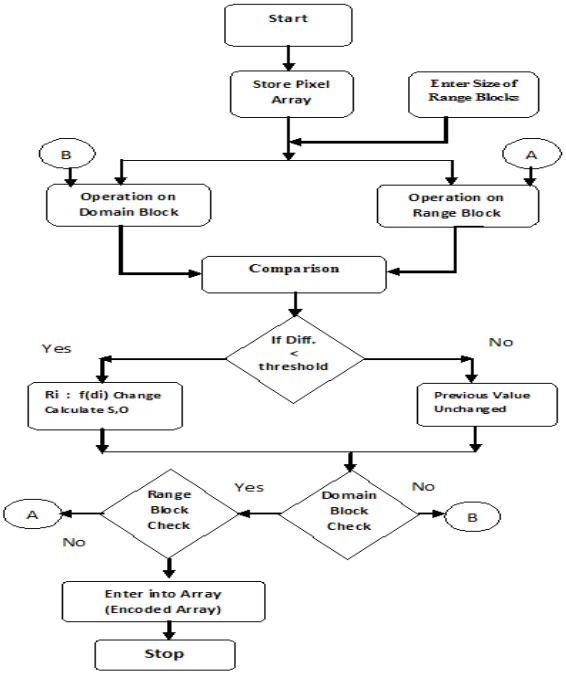


Fig. 13. Flowchart for Encoding Process

7 Conclusion

In proposed method, least sized range blocks provide higher PSNR at the cost of lesser compression ratio. Reduces the complexity of computation of encoding phase due to less number of domain blocks (Reduced by 25%). Encoding time is less. There is loss of data but higher compression rate is achieved. To the best of our knowledge, ours is the first technique to use Feature Extraction in Fractal compression. Thereby speeding up the encoding and decoding time.

Acknowledgements. We acknowledge Prof. M.L.Dhore, Prof. D.P Pawar for guiding us throughout the paper.

References

[1] Barnsley, M.: Fractals Everywhere. Morgan Kaufmann (1988)
[2] Jacquin, A.E.: A novel fractal block-coding technique for digital images. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1990, vol. 4, pp. 2225–2228 (April 1990)
[3] Chaurasia, V., Somkuwar: Speed Up Technique for Fractal Image Compression. In: IEEE International Conference on Digital Image Processing (ICDIP), pp. 319–323 (March 2009)

[4] Chaurasia, V., Somkuwar, A.: Improved Suitable Domain Search for Fractal Image Encoding. International Journal of Electronic Engineering Research, 1–8 (2010)
[5] Koli, N.A., Ali, M.S.: A Survey on Fractal Image Compression Key Issues. Information Technology Journal 7(8), 1085–1095 (2008)
[6] Jacquin, A.E.: Fractal image coding based on a theory of iterated contractive image transformations. In: SPIE, vol. 1360, pp. 227–239 (1990)
[7] Gonzales, R.C., Woods, R.E.: Digital Image Processing
[8] Fisher, Y.: Fractal Image Compression: Theory and Application. Springer (1995)
[9] Weistead, S.: Fractal and Wavelet Image Compression Technique. PHI, India (2005)
[10] Nixon, M.S., Aguado, A.S.: Feature extraction and image processing, 2nd edn. Academic Press, Oxford (2002)
[11] Zhao, E., Liu, D.: "Fractal image compression methods: a review. In: Third International Conference on Information Technology and Applications, ICITA 2005, July 4-7, vol. 1, pp. 756–759 (2005)
[12] Wang, H.: Fast Image Fractal Compression with Graph-Based Image Segmentation Algorithm
[13] Jaquin, A.E.: Image coding based on a fractal theory of iterated contractive image transformation. IEEE Trans. on Image Processing 1(1) (January 1992)
[14] Jaquin, A.E.: Fractal image coding: A review. Proceeding of tile IEEE 81(10) (October 1993)
[15] Distasi, R., Nappi, M., Riccio, D.: A range/domain approximation error based approach for fractal image compression. IEEE Trans. Image Processing 15(1), 89–97 (2006)
[16] Wang, H.: Fast Image Fractal Compression with Graph-Based Image Segmentation Algorithm

Case Study of Failure Analysis Techniques for Safety Critical Systems

Aiswarya Sundararajan[1] and R. Selvarani[2]

[1] Student, M.S. Ramaiah Institute of Technology, Bangalore
aisrajan@gmail.com
[2] Professor, M.S. Ramaiah Institute of Technology, Bangalore
selvss@yahoo.com, selvarani.riic@gmail.com

Abstract. Safety critical systems are built upon complex software and are difficult to maintain. These systems must effectively deal with the defects identified by analyzing its failure in order to make the system free from hazards. Any chance of human injury or death can be avoided by thoroughly verifying the safety of critical software embedded in any safety system. In this paper, the analysis on different failure analysis techniques such as Failure Modes, Effects Analysis (FMEA), Failure Modes, Effects and Criticality Analysis (FMECA) and Fault Tree Analysis (FTA) are carried out considering dependability as its critical parameter. The risk involved in safety critical system is analyzed with the case study of remote monitoring of a patient with pacemaker. The main observations are: i) Failure mode classification of the software at every stage, ii) Safety critical parameter evaluation, iii) Indication of defensive measures against the severity of hazards, iv) Correlation of FMEA, FMECA and FTA with the computed critical data and v) Recommendation of an appropriate failure analysis method for pacemaker operation to ensure safety.

Index Terms— cardiac arrhymias, dependability, FMEA, FMECA, FTA, pacemaker, remote monitoring and safety critical systems.

1 Introduction

The failure in the context of a safety critical system can be defined per [1] as “the non performance or inability of the system or component to perform its intended function for a specific time under specified environmental conditions." Thus the software is said to be safe, if it is impossible to produce an output that could cause an undesired event for the system [2]. The so called patient remote monitoring for pacemaker operation is a big step towards continuous monitoring of the victim from home, making it possible for the physician to access the data from anywhere [3]. Due to its safety criticality, a dependability study is initiated that emphasises on the level of user’s trust and confidence in operating the system up to an expected level [4]. Some of the popular software failure analysis techniques those are relevant to the pacemaker software is examined in this paper with the notion of reducing the hazards. Our contributions are three folds. First, the pacemaker software is analyzed for all possible failure modes by considering its critical parameters based upon the dependability study [5]. Second, the estimation of these parameters is accomplished by adopting FMEA, FMECA and FTA

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 367–377.
springerlink.com

and also the preventative methods are proposed. Finally, a well-suited analysis method for pacemaker operation is intimated. The remaining of this paper is organized as follows. The current state-of-art in the verification of safety critical system is described in Section 2. The characterization of pacemaker software is examined in Section 3. The software faults are recognized in section 4. The result of the failure analysis is presented in Section 5. An advanced failure analysis tool, CARA FAULT TREE is used to analyze the effectiveness of FTA in our study. A comparative study of the techniques FMEA, FMECA and FTA are represented in section 6. Then the work is concluded, and the future scope of research direction is revealed.

2 Background and Related Work

"A set of software operations if not performed, performed out-of-sequence or performed incorrectly might result in hazardous condition" [2][7] is called safety critical software. Such software must follow the below mentioned trademarks: - controlling safety critical hardware, monitoring state of the system, sensing of alarming conditions and display information, handling of fault detection and generating status of safety critical hardware. The above functions are stated in [6][7]. The safety critical software could become uncertain on account of following arguments cited [6][7]. Failure of software to perform a required function, software performs a function that is not required; software possesses timing and sequence problems, failure of software to recognize a hazardous condition and failure of software to pinpoint a safety critical function. Failure analysis is the method of reducing the system hazards and failure modes, then determine which of those are caused by or influenced by software or lack of software. The failure analysis is complicated by several dependability benchmarks as shown in the table 1.

Table 1. Dependability Attributes

Dependability Factors	Criteria
Availability	Mean Time To Failure (MTTF), Mean Time To Repair (MTTR)
Reliability	Survival probability, Rate of failure of occurrence
Safety	Severity class, Rate of detecting failure and Risk Priority Number (RPN)

All these factors are determined, and hazardous failure condition at the system level is averted. The failure analysis of pacemaker software is investigated using failure analysis techniques is the related work. Both FMEA and FMECA are bottom-up approaches [9]. They are structured, table based process of exploring the ways in which a software component can fail and its consequences [9]. But FTA is a top-down approach that uses Boolean logic to form a fault tree structure with an undesired event (failure) called TOP as the root. Multiple failures and repairs in the safety critical system are analyzed through an effective tool, CARA FAULT TREE.

3 Case Study of Remote Monitoring of a Pacemaker

One of the most widely used implantable devices, artificial pacemakers, are useful in treating abnormal heart beats [10]. The overview of pacemaker system is given below:

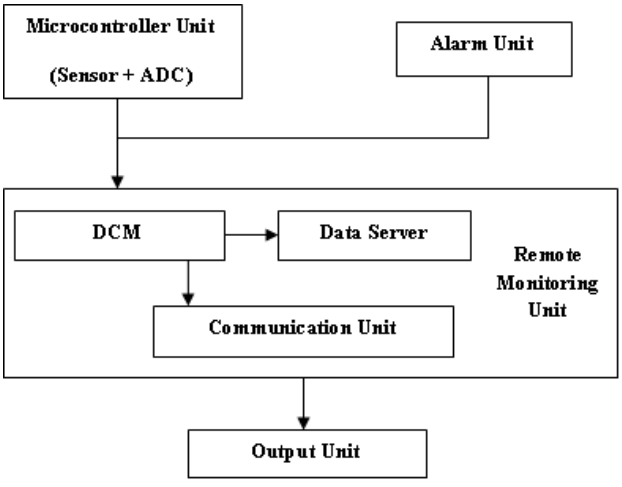


Fig. 1. Block diagram of pacemaker

The ECG (electrocardiographic) recording of patients is continuously done with suspected cardiac arrhymias [10] and is accomplished in the remote monitoring of the pacemaker. The wireless automatic transmissions of the abnormal ECG waveform signals are sent from patient's home to the monitoring center, and thereby the medical practitioner can modify the device settings without any surgery [10][11]. The component description of the pacemaker system is summarized in the table 2.

Table 2. Pacemaker Components

Components	Functions
Microcontroller unit	Captures signal from the implant and sends digital signal to remote monitoring unit through ADC
Remote Monitoring Unit	Recorded patient data is displayed on DCM and is sent to communication unit through data server (stores the history of patient records)
Communication Unit	Consists of an inbuilt GSM/Bluetooth/Radio frequency module for communicating with the recipient
Output Unit	Data is received as email/fax/sms to the physician
Alarm Unit	Activated when analog voltage is greater than the specified threshold and a message is sent to the physician

4 Critical Parameters of Dependability

Dependability is described as "the system's characteristic that justifies placing one's reliance on it" [12]. In order to rely on the system, it needs to satisfy certain attributes like availability, safety, reliability and security as mentioned in [13] and their parameters are summarized below.

Table 3. Critical dependability parameters

Dependability Measures	Description
MTTF	Mean time to first system failure
System Availability	Average system availability in time t
Survival Probability R_0	Probability that an undesired event does not occur at time t
Unavailability Q_0	Probability that an undesired event occurs at time t
MTTR	Mean time to repair
Severity Ranking (S)	Rates the high risk of potential effect of the failure
Rate of detecting failure (D)	Probability that the problem can be detected before it reaches the end user
Rate of occurrence of failure (O)	Number of times a failure mode occurs
RPN	Threshold value that computes the product of S, D and O

5 Failure Analysis

The cause of any failure must be determined if a component or product fails in service, the corrective actions can be taken to eliminate or control the risk [12] [13] based on the severity of its effects. FMEA, FMECA and FTA are significant failure analysis techniques that are featured below. FMEA analyses different failure and their effects on the system [14]. Troubleshooting of the system and their corrective actions is rendered through the iterative process. The illustration of FMEA is given at the end of the paper (see table 5). The FMEA table depicts the failure modes of function/system or subsystem in an individual component or program module, failure detection methods, their compensating provisions, etc. The probability of failure can be distinguished by the severity class ranking [14] [15]. FMECA is a separate analysis of FMEA and criticality and criticality pertains to the measure of probability of occurrence of a failure mode. FMECA is highlighted by the rate of severity, occurrence, detection and as well as RPN values [15]. The calculation of RPN is derived from the following equation,

$$\text{Risk Priority Number, RPN} = \text{Severity} \times \text{Occurrence} \times \text{Detection} \quad (1)$$

This equation is witnessed by lower detection (less chance for the failure mode to escape detection), severity and occurrence values for hazardless operation of the system. The counteracting alertness is urged on the occasion of the highest RPN values. The typical criticality parameters are disclosed at the end of the paper in the table 5. Two graphs are depicted (figure 2 and 3) to accommodate FMEA and FMECA, with the first one being occurrence/severity graph and the second is the failure mode/RPN value graph. They provide an additional way to use the rating scales to prioritize potential problems [15].

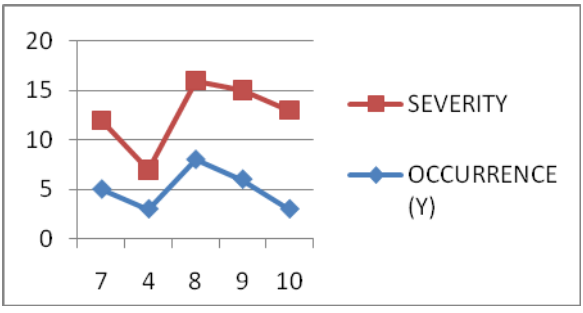


Fig. 2. Occurrence Vs Severity

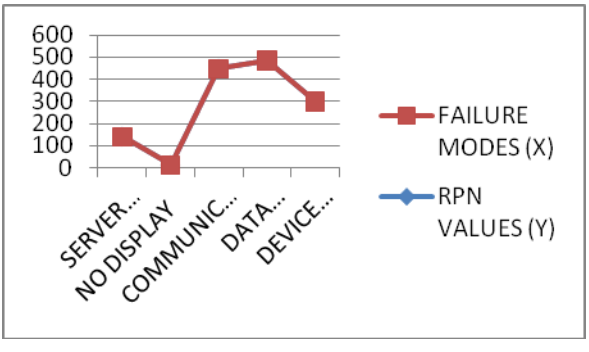


Fig. 3. Failure modes Vs RPN values

Fault tree analysis (FTA) probes the possible software causes of potential pacemaker component failures. The common failures can be generalized by constructing the top-level FTA tree [16]. Each of the selected failures becomes the TOP undesired event (root). The analysis proceeds by determining how the TOP event can be caused by individual or combination of lower level failures or events. Higher gates are the outputs from lower gates in the tree. Top event is the output of all the input faults or events that occur [16] [17]. The qualitative and quantitative analyses are the two divisions in FTA evaluation. Cut sets are determined in qualitative assessment. They are the set of event combinations that can cause a TOP event to occur. The dependability measures Mean Time to Failure (MTTF), Mean Time to Repair (MTTR), reliability and availability [18] are measured. The top-down construction of fault trees can be composed using CARA-Fault Tree software tool. The current version of CARA-Fault Tree is v4.1. Both qualitative and quantitative analyses are derived from the CARA tool. And also the correctness and consistency checking of the fault tree can be done in case of any illegal couplings [19] [20]. The fault trees of each failure

mode are drawn along with the failure data with the CARA tool [13] [21]. Some of the snapshots of failure modes (see figures 4 and 5) using this tool are shown below:

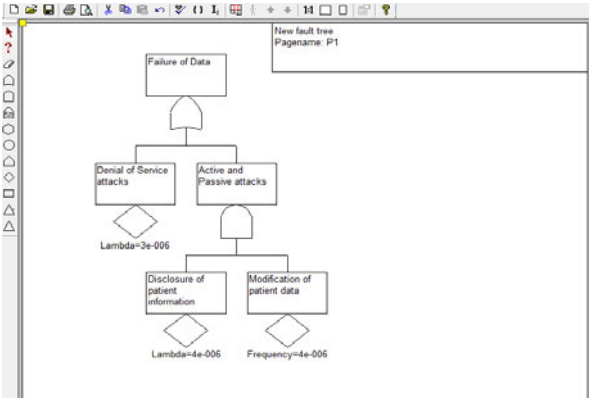


Fig. 4. Data Failure

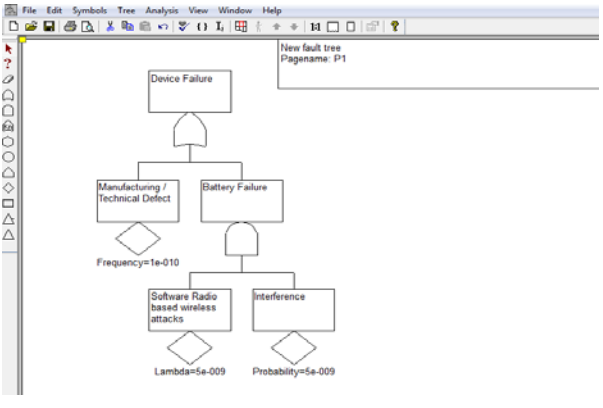


Fig. 5. Device Failure

The detailed FTA evaluation is interpreted at the end of the paper (see table 6). From the FTA worksheet, reviews figured out are summarized as follows. For the lesser value of MTTF for a TOP event, the probability of its occurrence is more when comparing to other events. For instance, MTTF is lower for DCM failure. Then if $A_{0,}$ av (t) increases, availability of that particular component is reduced. The final one is that the probability of occurrence of the TOP event never increases until FREQ value is escalated. A graph is drawn by with the assistance of failure modes in horizontal and MTTF in the vertical axis (see figure 6).

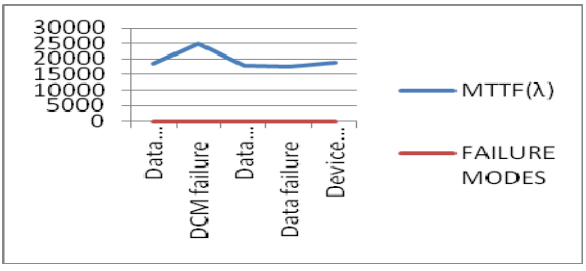


Fig. 6. MTTF vs Failure modes

6 Comparison of FMEA and FMECA vs. FTA

Of the three failure analysis methods that are addressed, FTA is found to be more powerful in terms of determining the probability of the failure rates and occurrence of the failure and component importance. The effectiveness of the failure analysis can be increased by combining FMECA and FTA in the design process [18]. The advantages of FTA over FMEA and FMECA is summarized in the table 5 and thereby justified that FTA is a forceful method for our case study of the pacemaker. The comparison table is given below:

Table 4. Technical comparison of FMEA, FMECA and FTA

Factors	FMEA	FMECA	FTA
Tree structured picturisation	NO	NO	YES
Failure Mode Identification	YES	YES	YES
Failure Detection	YES	YES	NO
Criticality Analysis	NO	YES	YES
Probability of failure occurrence	NO	NO	YES
Probability of survivability of components	NO	NO	YES
Probability of system availability	NO	NO	YES
Identifying the sub component failures that lead to TOP event	NO	NO	YES
Usefulness at multiple failures	NO	NO	YES
Probability of the component availability	NO	NO	YES
Suitability for large projects	YES	YES	NO

Table 5. FMEA and FMECA Criticality Analysis

Failure Modes	FMECA				FMEA	
	Severity Rating (S)	Occurrence Rating (O)	Detection Rating (D)	RPN (S*O*D)	Compensating Provision	Severity Class
Data Server	7 (High)	5(Moderate)	4 (Moderate)	140	Backup server (proxy server)to take the role of the original server and thereby achieving reliability	7 (High)
DCM	4 (Low to Moderate)	3 (Low)	1 (Almost certain)	12	Warning alarm for the connection loss and problem with the lead wires	4 (Low to Moderate)
Data Transfer	8 (Very High)	8 (High)	7 (Very Low)	448	Good design of communication protocol and achieving location transparency	8 (Very High)
Data Handling	9 (Hazard)	6 (Moderate to High)	9 (Very Remote)	486	-Enhance the existing security algorithms -Implementation of wireless firewall -Can create a secure channel for communication of the patient data	9 (Hazard)
Pacemaker – The device	10 (Hazard)	3 (Low)	10 (Almost Impossible	300	-Monitoring of access points to avoid wireless tapping	10 (Hazard)

Table 6. FTA detailed evaluation Worksheet

<table>
<tr><th>Failure Modes</th><th>cut set</th><th>$Q_0(t)$</th><th>$A_0(t)$</th><th>$R_0(t)$</th><th>MTTF λ</th><th>FREQ (TOP) Occurrence per hr/10^6 hr</th><th>Cut set importance</th></tr>
<tr><td>Data server</td><td>2</td><td>t = 876
<table><tr><th>t</th><th>Q0(t)</th></tr><tr><td>0</td><td>0.0000e+000</td></tr><tr><td>876</td><td>8.7861e-004</td></tr><tr><td>1752</td><td>1.7535e-003</td></tr><tr><td>2628</td><td>2.6275e-003</td></tr><tr><td>3504</td><td>3.5009e-003</td></tr><tr><td>4380</td><td>4.37 4e-003</td></tr><tr><td>5 56</td><td>5.2452e-003</td></tr><tr><td>6132</td><td>6.1162e-003</td></tr><tr><td>7008</td><td>6.9865e-003</td></tr><tr><td>7884</td><td>7.8560e-003</td></tr><tr><td>8760</td><td>8.7247e-003</td></tr></table></td><td>0.9905</td><td>t =550
<table><tr><th>Time</th><th>Ro(t)</th></tr><tr><td>0</td><td>1.0000e+000</td></tr><tr><td>550</td><td>9.9890e-001</td></tr><tr><td>1154</td><td>9.9769e-001</td></tr><tr><td>1819</td><td>9.9637e-001</td></tr><tr><td>2551</td><td>9.9491e-001</td></tr><tr><td>3356</td><td>9.9331e-001</td></tr><tr><td>4241</td><td>9.9155e-001</td></tr><tr><td>5215</td><td>9.8962e-001</td></tr><tr><td>6286</td><td>9.8751e-001</td></tr><tr><td>7464</td><td>9.8518e-001</td></tr></table></td><td>18430

MTTR (¥)=3</td><td>2e-006</td><td>DOS attack with higher failure rate of

9.9966e-001</td></tr>
<tr><td>DCM</td><td>2</td><td>NA</td><td>0.6667</td><td>NA</td><td>248.43 MTTR (¥)=1</td><td>Negligible occurrence</td><td>Negligible occurrence</td></tr>
<tr><td>Data Transfer</td><td>2</td><td>t=876
<table><tr><th>t</th><th>Value</th></tr><tr><td>0</td><td>0.0000e+000</td></tr><tr><td>876</td><td>4.3704e-003</td></tr><tr><td>1752</td><td>8.7217e-003</td></tr><tr><td>2628</td><td>1.3054e-002</td></tr><tr><td>3504</td><td>1.7367e-002</td></tr><tr><td>4380</td><td>2.1662e-002</td></tr><tr><td>5256</td><td>2.5938e-002</td></tr><tr><td>6132</td><td>3.0195e-002</td></tr><tr><td>7008</td><td>3.4433e-002</td></tr><tr><td>7884</td><td>3.8653e-002</td></tr><tr><td>8760</td><td>4.2855e-002</td></tr></table></td><td>0.9768</td><td>t=550
<table><tr><th>Time</th><th>Ro(t)</th></tr><tr><td>0</td><td>1.0000e+000</td></tr><tr><td>550</td><td>9.9725e-001</td></tr><tr><td>1154</td><td>9.9424e-001</td></tr><tr><td>1819</td><td>9.9094e-001</td></tr><tr><td>2551</td><td>9.8732e-001</td></tr><tr><td>3356</td><td>9.8336e-001</td></tr><tr><td>4241</td><td>9.7901e-001</td></tr><tr><td>5215</td><td>9.7425e-001</td></tr><tr><td>6286</td><td>9.6905e-001</td></tr><tr><td>7464</td><td>9.6335e-001</td></tr><tr><td>8760</td><td>9.5713e-001</td></tr></table></td><td>17923</td><td>5.002e-006</td><td>Failure of Network connectivity with failure rate of 1.0000e+000</td></tr>
<tr><td>Data Handling</td><td>4</td><td>t=876
<table><tr><th>t</th><th>Value</th></tr><tr><td>0</td><td>0.0000e+000</td></tr><tr><td>876</td><td>2.6245e-003</td></tr><tr><td>1752</td><td>5.2422e-003</td></tr><tr><td>2628</td><td>7.8530e-003</td></tr><tr><td>3504</td><td>1.0457e-002</td></tr><tr><td>4380</td><td>1.3054e-002</td></tr><tr><td>5256</td><td>1.5644e-002</td></tr><tr><td>6132</td><td>1.8228e-002</td></tr><tr><td>7008</td><td>2.0805e-002</td></tr><tr><td>7884</td><td>2.3374e-002</td></tr><tr><td>8760</td><td>2.5938e-002</td></tr></table></td><td>0.995709</td><td>t= 550
<table><tr><th>Time</th><th>Ro(t)</th></tr><tr><td>0</td><td>1.0000e+000</td></tr><tr><td>550</td><td>9.9726e-001</td></tr><tr><td>1154</td><td>9.9304e-001</td></tr><tr><td>1819</td><td>9.8833e-001</td></tr><tr><td>2551</td><td>9.8328e-001</td></tr><tr><td>3356</td><td>9.7776e-001</td></tr><tr><td>4241</td><td>9.7172e-001</td></tr><tr><td>5215</td><td>9.6511e-001</td></tr><tr><td>6286</td><td>9.5791e-001</td></tr><tr><td>7464</td><td>9.5004e-001</td></tr><tr><td>8760</td><td>9.4146e-001</td></tr></table></td><td>17613.4</td><td>3.13773e-006</td><td>DOS attacks and disclosure of patient information are to be given critical importance based on the failure rates of 1.0000e+000</td></tr>
<tr><td>Device-Pacemaker</td><td>4</td><td>NA</td><td>1</td><td>NA</td><td>18777</td><td>1.00219e-010</td><td>Software radio based wireless attacks to be given more importance based on the failure rate of 9.9976e-001</td></tr>
</table>

7 Conclusions

In the present scenario, safety critical systems are more pervasive in the field of medicine and are designed with utmost care. The dependability requirement is an important criterion in these systems. To reduce the probability of losses, appropriate failure analysis methodologies are used to verify the safety of the critical system. The reason for the failure of the critical components as well as the precautionary measures to avoid any potential risks is shown. The experimental results showed that the computed criticality parameters can increase the assurance of dependability level of the pacemaker system. FTA seemed to be very effective in terms of analyzing accurate failure modes of the pacemaker system. Our research will continue by analyzing the software of a safety critical system to measure the safety parameters of the same at the design level.

Acknowledgments. I would like to express my gratitude to my professor, Dr. R. Selvarani (Head of the department, MSRIT, Bangalore) for her kind co-operation, support and encouragement which helped me in completion of this paper.

References

[1] Greenwell, W.S., Knight, J.C.: Framing analysis of software failure with safety cases. IEEE Transactions on Software Engineering 22904 (January 2010)
[2] Kornecki, A.J.: Assesment of software safety via catastrophic events coverage. In: Proceedings of IEEE Computer Society (March 2007)
[3] Muller, A., Helms, T.M., Wildau, H.-J.: Remote Monitoring in Patients with Pacemakers and Implantable Cardioverter-Defibrillators: New Perspectives for Complex Therapeutic Management. University of Heidelberg, Germany (2009)
[4] Mauser, H., Thurner, E.: Electronic Throttle Control-A Dependability case study. Jounal of Universal Computer Science 5(10), 730–741 (2009), Siemens, A.G.
[5] SoftWcare, S.L., Serafin Avendano, C.: Safety and dependability analysis to complement testing of safety-critical software. Espana (2009)
[6] Ben Swarup, M., Seetha Ramaiah, P.: An approach to modeling software safety, department of computer science and systems engineering. In: 9th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing. IEEE Computer Society Press (2010)
[7] Ben Swarup, M., Seetha Ramaiah, P.: A software safety model for safety critical applications, department of computer science and systems engineering, Andhra University. International Journal of Software Engineering and its Applications 3(4) (October 2009)
[8] Glancey, J.: Failure Analysis Methods- What, Why and How? Special Topics in Design. MEEG 466 (Spring 2006)
[9] FMECA for command, control, communications, computer, intelligence, surveillance and reconnaissance (C4ISR). Technical Manual (September 29, 2006)
[10] Halperin, D., Heydt-Benjamin, T.S., Ransford, B., Clark, S.S., Defend, B., Fu, K., Kohno, T., Maisel, W.H.: Pacemakers and Implantable Cardiac Defibrillators:Software Radio Attacks and Zero-Power Defenses (2010)
[11] Sharma, A.: Towards A Verified Cardiac Pacemaker. A Technical Report (November 2010)

[12] Barbacci, M., Klein, M.H., Longstaff, T.H., Weinstock, C.B.: Principles for Evaluating the Quality Attributes of a Software Architecture. SEI, Carnegie Mellon University (March 1997)
[13] User's Manual for CARA-FaultTree v4.1 by Sydvest Software
[14] Crowe, D., Feinberg, A.: Design for Reliability. In: Failure Modes and Effects Analysis, ch. 12. CRC Press, Bocaraton (2001), http://www.reliasoft.com/newsletter/2q2003/rpns.htm
[15] Assessment worksheet of a particular piece of equipment for UMS managers,supervisors,OHS consultation committees and representatives for all complex asessments, Hazard Identification. Risk Assessment and Control Procedure. University of Western Sydney, June 23 (2003)
[16] Reliability engineering resources-Fault Tree Handbook (NUREG-0492), US Nuclear Regulatory Commission, http://www.weibull.com/basics/faulttree/index.htm
[17] Carlo KoppPeter Harding & Associates, Pty Ltd., System Reliability and Metrics of Reliability, Copyright, PHA Pty Ltd. (1996), http://www.pha.com.au/
[18] FAULT TREE ANALYSIS- A Special Bibliography from the NASA Scientific and Technical Information (STI) Program, http://www.sti.nasa.gov/new/fta34.pdf
[19] Sommerville, I.: Software Engineering. In: Insulin Pump, 7th edn., vol. ch. 3 (2009)
[20] Otto, K.W., Kristin: Product Design - Techniques in Reverse Engineering and New Product Development. Prentice Hall (2001) ISBN 0-13-021271-7
[21] Kmenta, S., Ishii, K.: Scenario-Based Failure Modes and Effects Analysis Using Expected Cost. Journal of Mechanical Design- 126(6), 1027 (2004)

Implementation of Framework for Semantic Annotation of Geospatial Data

Preetam Naik[1], Madhuri Rao[2], S.S. Mantha[3], and J.A. Gokhale[4]

[1] M.E. Student, TSEC Bandra(W), Mumbai, India
preetget@yahoo.com
[2] Assistant professor, Bandra(W), Mumbai, India
my_rao@yahoo.com
[3] Professor, VJTI, Mumbai, India
ssmantha@vjti.orgin
[4] Assistant professor, VESIT, Mumbai, India
gokhalej@yahoo.com

Part I

Abstract. Framework is used to provide effective development, exchange, and use of geospatial as well as non-geospatial data. In previous paper, design of framework for semantic annotation of geospatial data, the architecture of framework and its services has been discussed briefly. This paper is based on the detailed explanation of implementation of some of the services provided by the framework.

1 Introduction

Geospatial data is basically used to identify the natural or constructed geographic location features on the Earth. On web, geospatial data is available in different geographic formats like remote sensing images, maps, textual data files etc. The retrieval of these data requires special attention due to geographical distribution of the sources and the heterogeneity of the data [1]. Instead of using traditional information retrieval techniques based on indexing and string matching, semantic annotation is used to specify the vocabulary of content's meaning. Semantic annotations are linked to ontologies and support logic-based reasoning. So, the search for a geospatial information retrieval framework becomes necessary which relies on ontologies, allowing users to retrieve desired data, based on their semantics [2].

Part II

Abstract. This chapter gives the detailed design of a framework. It explains the working of several services used in the framework.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 379–384.
springerlink.com

2 Background

The design of proposed framework used for semantic annotation of geospatial data is basically made up of different types of layers: client layer, service class layer, service layer and data layer. The architecture is shown in Figure 1.

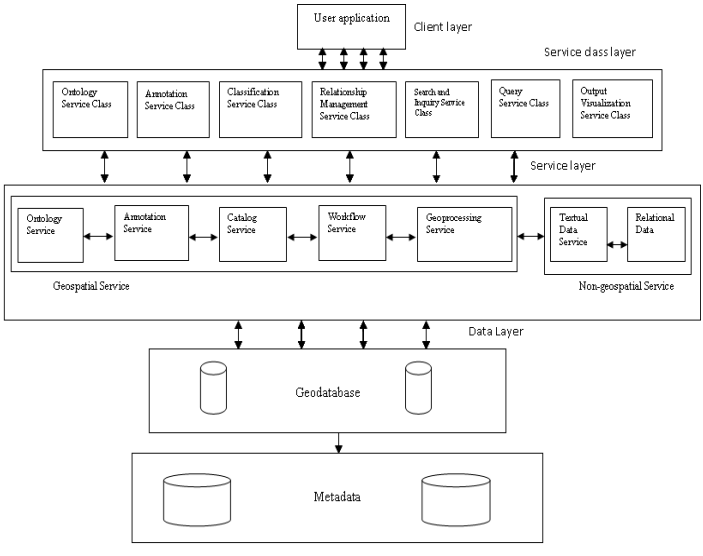


Fig. 1. Design of proposed framework

Client layer is responsible for communication between user and different services. Service Class layer provides several necessary services such as high performance GIS services, multi-access to other services, high stability and reliability (e.g. load balance) and high security. It provides seven types of service classes as: ontology service class, annotation service class, classification service class, relationship management service class, search and inquiry service class, query service class and output visualization service class. Service layer provides different value-added services to its upper layer. They are: ontology service, annotation service, catalog service, workflow service and geoprocessing service [1].

Geodatabase is used for storing, indexing, querying, and manipulating geographic information and spatial data. Metadata provides content, quality, type, creation, and spatial information about a data set.

Part III

Abstract. This chapter explains the detailed implementation of some of the services like Ontology service, Annotation service, Query Service, Search and Inquiry service and Output Visualization service.

3 Service Classes

3.1 Ontology-Based Service

Ontology service is responsible for handling ontologies. It provides wide range of operations to store, manage, search, rank, analyze and integrate ontologies. Generally geospatial services have no semantic service description. Ontologies are used within the context of geospatial data infrastructures to denote a formally represented knowledge that is used to improve data sharing and information retrieval. So it is very difficult and time consuming to invoke a geospatial service correctly [2]. To solve this problem, ontology is explicitly used for semantic service description. The overall framework of an ontology service is presented in figure 2 as:

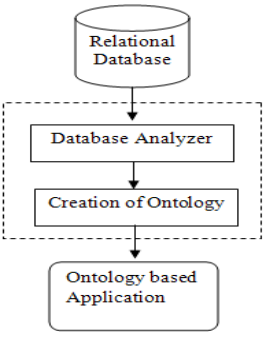


Fig. 2. Framework of ontology service

The input for the framework is data stored in relational database. The framework uses the database analyzer to extract schema information from database such as, the primary keys, foreign keys and dependencies. Using obtained information ontology is created. The frame is domain/application independent and can create ontologies for general and specific domains from relational database. The snapshot of an ontology service is shown in the following Figure 3.

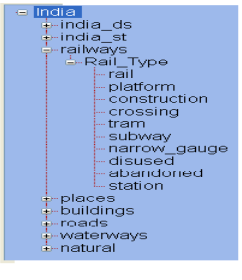


Fig. 3. Snapshot of ontology service

3.2 Annotation-Based Service

Annotation service semantically annotates different kind of geospatial data, such as satellite images, maps and graphs [2]. In geographic applications, annotations should also consider the spatial component, since geographic information associates objects and events to localities, through places and geographic object names, spatial relationships and standards. Hence, the geospatial annotation process should be based on geospatial evidences – those that conduct to a geographic locality or phenomenon. The workflow of an annotation service is shown in the following Figure 4.

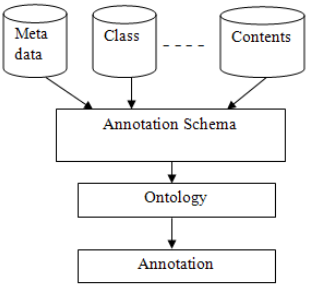


Fig. 4. Workflow of annotation service

In this service, annotation is generated with the help of various sources like metadata, different classes, contents etc. First the annotation schema is defined then the schema is filled with ontology terms. Then the framework has to relate them with a semantic meaning for annotation creation.

After implementation of an annotation service, the snapshot is shown in the following Figure 5. In this, the annotation of the selected region is displayed. An id and the name of state are displayed.

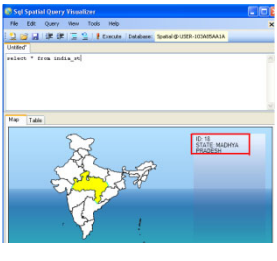


Fig. 5. Snapshot of annotation service

3.3 Query Service

The Query service is implemented which allows selecting and displaying attribute records of interest in grid format. The flow of query service is shown in following Figure 6.

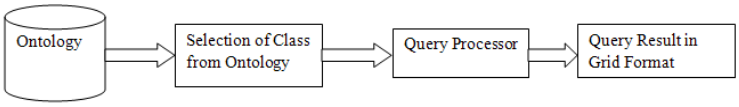


Fig. 6. Flow of query service

The snapshot of query service after implementation is given in the following Figure 7.

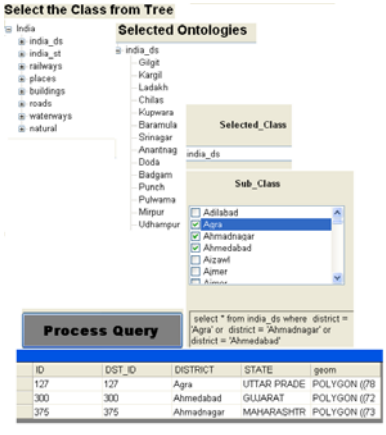


Fig. 7. Snapshot of query service

In this way first the class is selected. Then ontology is displayed regarding to the class. Also the subclasses are displayed. After selecting the process query option, query is processed on the subclasses which are selected and the final result is produced in grid format.

3.4 Search and Inquiry Service

This service allows searching and inquiring data from the service layer. Usually, the search for these data and methods is done by their syntactic content, focusing primarily in keyword matching. This can lead to the retrieval of irrelevant data, disregarding relevant files. Hence, semantic interoperability is a key issue in discovery, access and effective search for data in different application contexts.

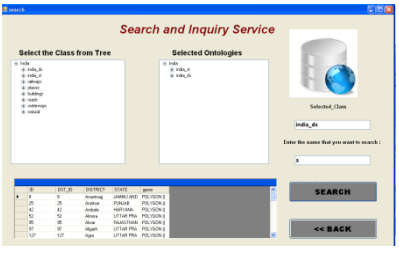


Fig. 8. Snapshot of Search and Inquiry service

3.5 Output Visualization Service

This service converts vector file into raster file means .shp file is converted into grid format as well as it shows results on the map. The query is executed and the result is displayed on the map which is shown in Figure 9.

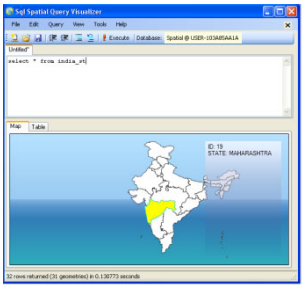


Fig. 9. Snapshot of output visualization service on map

Acknowledgments. I would like to express a deep sense of gratitude towards my guide Ms. Madhuri Rao for her constant encouragement, intellectual and valuable suggestions. The work that I am able to present would just not have been possible without her timely guidance.

I would also like to thank my family for providing their support during project work.

I am also thankful to all staff members for all the help and co-operation they have rendered to me all the time.

Ms. Preetam Naik.

References

1. Naik, P., Rao, M., Mantha, S.S., Gokhale, J.A.: Design of Framework for Semantic Annotation of Geospatial Data. In: IEEE International Conference on Network Communication and Computer, pp. 282–285 (March 2011)
2. Marcario, C.G.N., Medeiros, C.B.: Specification of a framework for semantic annotation of geospatial data on the web. In: IEEE International Conference on Challenges in Environmental Science and Computer Engineering, pp. 27–32 (March 2009)

Optimized Large Margin Classifier Based on Perceptron

Hemant Panwar and Surendra Gupta

Computer Engineering Department
Shree Govindram Seksariya Institute of Technology and Science, Indore, India
hemant061018@gmail.com,
sgupta@sgsits.ac.in

Abstract. Larger margin of separating hyperplane reduces the chances of generalization error of classifier. The proposed linear classification algorithm implements classical perceptron algorithm with margin, to reduce generalized errors by maximizing margin of separating hyperplane. Algorithm shares the same update rule with the perceptron, to converge in a finite number of updates to solutions, possessing any desirable fraction of the margin. This solution is again optimized to get maximum possible margin. The algorithm takes advantage of data that are linearly separable. Experimental results show a noticeable increment in margin. Some preliminary experimental results are briefly discussed.

1 Introduction

In the field of machine learning, the goal of classification is to use an object's characteristics to identify which class (or group) it belongs to. An object's characteristics are also known as feature values and are typically presented to the machine in a vector, called a feature vector. A classifier separates classes using a separation boundary or hyperplane. Classifiers generally face the problem of incorrect classification for those instances which are closer to solution hyperplane, known as generalization error.

Study shows that generalization error can be reduced by maximizing the margin which is the distance between instances and separation boundary [1]. This justifies high interest in Support Vector Machines (SVMs) [2]. SVMs produce large margin solutions by solving a constrained quadratic optimization problem using dual variables. Quadratic dependence of their memory requirements in the number of training examples prohibits the processing of large datasets. To overcome this problem, decomposition methods [3, 4] were developed that apply optimization only to a subset of the training set by keeping fixed a large number of variables and optimizing with respect to the set of the remaining constraints, known as active or working set. These algorithms are based on the Sequential Minimal Optimization (SMO) algorithm [3], in which the size of the active set is fixed to 2. The number of kernel rows to be cached remains crucial, since the memory hit rate is a factor that can considerably affect the performance of an algorithm. Although such methods have led to improved convergence rates, but in practice their superlinear dependence on the number of examples, lead to excessive runtimes, when large datasets are processed.

The above considerations motivated research in alternative way for large margin classifiers. Such algorithms are mostly based on the perceptron [5, 6]. Likewise the

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 385–392.
springerlink.com

perceptron algorithm, they focused on the primal problem by updating a weight vector which represented current state of the algorithm, whenever a data point presented to it satisfies a specific condition. Such algorithms processed one example at a time which allowed them to spare time and memory resources for handling large datasets. Subsequently, various algorithms succeeded in attaining maximum margin approximately by employing modified perceptron like update rules. Such algorithms included ROMMA [7], ALMA [8].

2 Motivation of the Algorithm

Consider a linearly separable training set $\{(x_k, l_k)\}_{k=1}^{m}$, with vectors x_k as input samples vector and labels $l_k \in \{+1,-1\}$. An augmented space is constructed by placing x_k in the same position at a distance ρ in an additional dimension, i.e. extending x_k to $[x_k, \rho]$ [9]. It can be also refer as hyperplane possessing bias in the non-augmented feature space. Following the augmentation, a reflection is performed with respect to the origin of the negatively labeled patterns by multiplying every pattern with its label. This allows a uniform treatment of both categories of patterns. So, $R \equiv \max_k \|y_k\|$ with $y_k \equiv [l_k x_k, l_k \rho]$ which represent the k^{th} augmented and reflected pattern. Obviously, $R \geq \rho$.

The relation characterizing optimally correct classification of the training patterns y_k by a weight vector u of unit norm in the augmented space is

$$u \cdot y_k \geq \gamma_d \equiv \max_{u':\|u'\|=1} \min_i \{u', y_i\} \forall k \tag{1}$$

where γ_d is the maximum directional margin. In proposed algorithm the augmented weight vector a_t is initially set to zero, i.e. $a_0 = 0$, and is updated according to the classical perceptron rule

$$a_{t+1} = a_t + y_k \tag{2}$$

each time an appropriate misclassification condition is satisfied by a training pattern y_k. Inner product of (2) with the optimal direction u and (1) gives

$$u \cdot a_{t+1} - u \cdot a_t = u \cdot y_k \geq \gamma_d$$

a repeated application of which gives [6]

$$\|a_t\| \geq u \cdot a_t \geq \gamma_d t$$

thus an upper bound can be obtain on γ_d provided $t > 0$

$$\gamma_d \leq \frac{\|a_t\|}{t} \tag{3}$$

It would be very desirable that $\|a_t\|/t$ approaches γ_d with t increasing since this would provide an after-run estimate of the accuracy achieved by an algorithm employing the classical perceptron update.

Assume that satisfaction of the misclassification condition by a pattern y_k has as a consequence that $\|a_t\|^2/t > a_t \cdot y_k$ (i.e., the normalized margin $u \cdot y_k$ of y_k (with $u_t \equiv a_t/\|a_t\|$) is smaller than the upper bound (3) on γ_d). Statistically, at least in the early stages of the algorithm, most updates do not lead to correctly classified patterns (i.e., patterns which violate the misclassification condition) and as a consequence $\|a_t\|/t$ will have the tendency to decrease. Obviously, the rate at which this will take place depends on the size of the difference $\|a_t\|^2/t - a_t \cdot y_k$ which, in turn, depends on the misclassification condition.

For solutions possessing margin the most natural choice of misclassification condition is the fixed (normalized) margin condition

$$a_t \cdot y_k \le (1-\varepsilon)\gamma_d \|a_t\| \tag{4}$$

with the accuracy parameter ε satisfying $0 < \varepsilon \le 1$. This is an example of a misclassification condition which if it is satisfied ensures that $\|a_t\|^2/t > a_t \cdot y_k$. The perceptron algorithm with fixed margin condition converges in a finite number of updates to an ε-accurate approximation of the maximum directional margin hyperplane [10, 11].

The above difficulty associated with the fixed margin condition may be remedied if the unknown γ_d is replaced for $t > 0$ with its varying upper bound $\|a_t\|/t$ [12]

$$a_t \cdot y_k \le (1-\varepsilon)\frac{\|a_t\|^2}{t} \tag{5}$$

Condition (5) ensures that $\|a_t\|^2/t - a_t \cdot y_k \ge \varepsilon \|a_t\|^2/t > 0$. Thus, it can be expected that $\|a_t\|/t$ will eventually approach γ_d close enough, thereby allowing for convergence of the algorithm to an ε-accurate approximation of the maximum directional margin hyperplane. It is also apparent that the decrease of $\|a_t\|/t$ will be faster for larger values of ε.

The proposed algorithm, now employs the misclassification condition (5) (with its threshold set to 0 for $t = 0$), which may be regarded as originating from (4) with γ_d replaced for $t > 0$ by its dynamic upper bound $\|a_t\|/t$.

3 Proposed Algorithm

The algorithm employing the misclassification condition above will attain maximum margin, but it can be further optimize. Figure 1 presents a basic output of algorithm discussed in previous section. Each circle represents a sample data, filled circles belong to positive class while empty circles belong to negative class. Two lines are marginal boundaries and one middle line is optimum separation boundary between two classes. Notice that, negative class samples are closer to optimum separation boundary than positive class samples. As margin is distance between closest samples and separation boundary. It can be further optimize by moving optimum surface boundary towards positive class samples till the distance of optimum surface from both classes become equals, Figure 2.

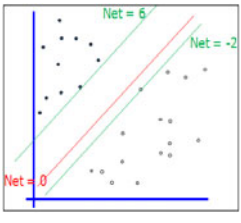


Fig. 1. Output of large margin classifier before optimization, separation boundary is closer to one particular class that leads to small margin. Moving the separation boundary can further increase the margin of classifier.

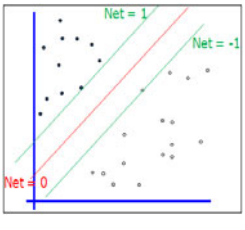


Fig. 2. Output of large margin classifier after optimization. In optimization the separation boundary is moved towards positive class such that distance of separation boundary from both classes becomes approximately equal. It increases the margin of separation boundary.

To get optimized margin, support vectors should be known. Support vectors can be identified by using any search algorithm i.e. linear search, by identifying two samples (one from each class) having lowest net value (b_t) from same class belonging samples. Let non reflected $x_{11}, x_{12}, \ldots, x_{1n}$ be the closest positive class support vector and $x_{21}, x_{22}, \ldots, x_{2n}$ be the closest negative class support vector and p_{t1}, p_{t2} be the net values of support vectors, respectively. To get an optimized margin, net value for support vector of positive class should be +1 and for support vector of negative class it should be -1.

Input : A linearly separable augmented dataset $S = (y_1, \ldots, y_k, \ldots, y_n)$ with reflection assumed.
Fix : ε
Define : $q_k = \|y_k\|^2$, $\varepsilon' = 1 - \varepsilon$
Initialize : $t = 0, a_0 = 0, l_0 = 0, \theta_0 = 0$;
repeat
  for $k = 0$ to n do
    $p_{t_k} = a_t \,.\, y_k$
    if $p_{t_k} \leq \theta_t$ then
      $a_{t+1} = a_t + y_k$
      $l_{t+1} = l_t + 2p_{t_k} + q_k$
      $t \leftarrow t + 1$
      $\theta_t = \varepsilon' l_t / t$
until no update made within the for loop
for $k = 0$ to n do
    $p_{s_1} = smallest\ in\ class + 1$
    $p_{s_2} = smallest\ in\ class - 1$
$\alpha = 2 / (p_{s_1} + p_{s_2})$
$\beta = (2a_n + p_{s_2} - p_{s_1}) / (p_{s_1} + p_{s_2})$
for only augmented weight in weight vector
  $a = \beta \,.\, a$
for other weights in weight vector
  $a = \alpha \,.\, a$

Fig. 3. Proposed algorithm for linear classification

To move position of separation boundary without changing its slope two parameters say α and β is used for new weights after optimization, α for weights other than augmented weight and β for augmented weight. For both support vectors equations are

$$\alpha\left(\sum_{i=0}^{n-1} a_i x_{1i}\right) + \beta\left(a_n x_{1n}\right) = +1 \tag{6}$$

$$\alpha\left(\sum_{i=0}^{n-1} a_i x_{2i}\right) + \beta\left(a_n x_{2n}\right) = -1 \tag{7}$$

simplifying (6),(7) for $x_{2n} = x_{1n} = 1$ (augmented dimension is always 1), $\sum_{i=0}^{n-1} a_i x_{1i} = p_{t1}$ and $-\sum_{i=0}^{n-1} a_i x_{2i} = p_{t2}$, will give

$$\alpha = \frac{2}{p_{t1} + p_{t2}}$$

and

$$\beta = \frac{2a_n + p_{t2} - p_{t1}}{a_n\left(p_{t1} + p_{t2}\right)}$$

these values of α and β can be used to get new values of weight vector as,

$$w_{new(0...n-1)} = a \times w_{old(0...n-1)}$$
$$w_{new(n)} = \beta \times w_{old(n)}$$

New value of weight vector will always attain an optimized margin. It provides a solution for optimization of surface boundary for linear classification. Figure 2 represents a proposed algorithm for linearly separable problem using misclassification condition (5) and above analysis.

4 Experimental Evaluations

The proposed algorithm is implemented in C++ using object oriented approach and applied on various linearly separable data sets. Results of these experiments are summarized in Table 1. Datasets used for training are Iris Plants Database (IRIS), Congressional Voting Records (VOTING), Ionosphere data set (IONOS) and Teaching Assistant Evaluation (TAE) data set. All of these data sets are obtainable from http://archive.ics.uci.edu/ml/datasets/. Experiments are performed on linearly separable subsets of these data sets. Data set files are modified into a system understandable file, for processing in the system. Proposed algorithm determines optimized weight vector value. Using this weight vector value margin is calculated as,

$$\gamma = \frac{\min_k \{a \cdot y_k\}}{\|a\|}$$

where γ is geometric margin i.e. directional margin γ_d normalized by weight vector (a).

Table 1. Results of Experiments with proposed algorithm

Data set	No. of Attributes	No. of Samples	Margin before Optimization (10^{-2})	Margin after Optimization (10^{-2})
IRIS	4	150	1.93441	41.7413
IONOS	36	310	0.0201952	0.0598356
VOTING	16	226	0.794151	0.794151
TAE	5	93	0.0698075	2.58847

Proposed algorithm has produced a noticeable increment in margin for IRIS, TAE and IONOS data sets. VOTING data set remains at approximately same margin before and after optimization, as VOTING data set has already attained maximum possible margin after learning. Above results showed that proposed algorithm always attains a maximum possible margin.

5 Conclusions

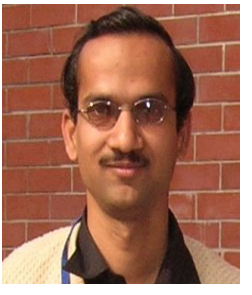
The proposed algorithm for large margin classifier based on perceptron employs the classical perceptron updates and converges in a finite numbers of steps. It uses required accuracy as only input parameter. Moreover, it is a strictly online algorithm in the sense that it decides whether to perform an update, taking into account only its current state and irrespective of whether the instance represented to it has been encountered before in the process of cycling repeatedly through the dataset. Optimization process further maximizes the margin.

References

[1] Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
[2] Cristianini, N., Shawe-Taylor, J.: An introduction to support vector machines. Cambridge University Press, Cambridge (2000)
[3] Platt, J.C.: Sequential minimal optimization: A fast algorithm for training Support vector machines. Microsoft Res. Redmond WA, Tech. Rep. MSR-TR-98-14 (1998)
[4] Joachims, T.: Making large-scale SVM learning practical. In: Advances in Kernel Methods Support Vector Learning. MIT Press (1999)
[5] Rosenblatt, F.: The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 65(6), 386–408 (1958)
[6] Novikoff, A.B.J.: On convergence proofs on perceptrons. In: Proc. Symp. Math. Theory Automata, vol. 12, pp. 615–622 (1962)
[7] Li, Y., Long, P.M.: The relaxed online maximum margin algorithm. Mach. Learn. 46(1-3), 361–387 (2002)
[8] Gentile, C.: A new approximate maximal margin classification algorithm. J. Mach. Learn. Res. 2, 213–242 (2001)
[9] Duda, R.O., Hart, P.E.: Pattern classification and scene analysis. Wiley (1973)
[10] Tsampouka, P., Shawe-Taylor, J.: Perceptron-like large margin classifiers. Tech. Rep., ECS, University of Southampton, UK (2005)
[11] Tsampouka, P., Shawe-Taylor, J.: Analysis of Generic Perceptron-Like Large Margin Classifiers. In: Gama, J., Camacho, R., Brazdil, P.B., Jorge, A.M., Torgo, L. (eds.) ECML 2005. LNCS (LNAI), vol. 3720, pp. 750–758. Springer, Heidelberg (2005)
[12] Panagiotakopoulos, C., Tsampouka, P.: The Perceptron with Dynamic Margin. In: Kivinen, J., Szepesvári, C., Ukkonen, E., Zeugmann, T. (eds.) ALT 2011. LNCS, vol. 6925, pp. 204–218. Springer, Heidelberg (2011)

Authors

Hemant Panwar received his Bachelor of Engineering degree in Information Technology from RGPV University, India in 2010. He is currently pursuing Master of Engineering in Computer Engineering from SGSITS, Indore, India. His research interests include machine learning and system software design.



Surendra Gupta received the Bachelor of Engineering degree in computer science and engineering from Barkatullah University, India in 1997 and Master of Engineering degree in computer engineering from DAVV University, India in 2000. He is currently working as Assistance Professors in computer engineering department at SGSITS Indore, India. His interests are in machine learning and optimization. He is a member of the computer society of India.

Study of Architectural Design Patterns in Concurrence with Analysis of Design Pattern in Safety Critical Systems

Feby A. Vinisha[1] and R. Selvarani[2]

[1] Department of ISE, AMC Engineering College, Bangalore, India
febyvinisha@gmail.com
[2] Department of CSE, M.S. Ramaiah Institute of Technology, Bangalore, India
selvss@yahoo.com

Abstract. Real time safety critical system is focused as it plays pivotal role in rendering software safety that strengthens hardware reliability to prevent hazardous failures. These problems can be addressed through the creation of quality design patterns which will effectively place the safety critical software parameter at the architectural level. This paper highlights the functionalities of each design pattern on the quantitative uphold on the safety quality factors. Here we worked on the design pattern that is compliant for safety critical systems pertaining to the safety parameters and quality attributes of various design patterns. Furthermore, we anticipate the additional quality attributes and features admissible to formulate this as the distinguished pattern for real time safety critical systems. The design patterns used in various applications is emphasized and characterized. More specifically the existing design patterns supporting safety critical systems is correlated to formulate the patterns based on the safety factors and quality attributes.

Keywords: Software Architecture, Design Patterns, Safety Critical Systems, Quality Attributes, Reliability, Robustness.

1 Introduction

Software architectures possess fundamental responsibility in the development of software systems and more specifically in the design phase. The basic definition of architecture holds true in software architecture also, as it frames a standard structure by exposing behavior of the system and hiding the implementation details. Based on the definition given by Grady Booch and redefined by Mary Shaw [1], Software Architecture encompasses the set of significant decisions about the organization of a software system including the selection of the structural elements and their interfaces by which the system is composed. According to the definition of Len Bass [2], software architecture of a program or computing system is the structure or structures of the system, which comprise software elements, the externally visible properties of those elements, and the relationships among them.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 393–402.
springerlink.com

The definition framed by Martin Fowler [3] says the software architecture is the highest-level breakdown of a system into its parts; the decisions that are hard to change; there are multiple architectures in a system; what is architecturally significant can change over a system's lifetime. IEEE 1471 defines architecture as the fundamental organization of a system embodied in its components, their relationships to each other and to the environment, and the principles guiding its design and evolution [4].

The advancement of design pattern aided designers and system architects to build alternate option of suitable solutions for commonly occurring recursive design problems and to compile safety critical systems. The distinguishing feature of safety critical systems is it generally involves monitoring or control of physical objects and operates in real time which may also be distributed across many operators and locations [7].According to the safety-critical standard IEC 61508, from a safety viewpoint, the software architecture is where the basic safety strategy is developed in the software [6]. Starting from basic web application to programming paradigm, now design pattern is widely used in Safety Critical Systems. Design patterns acts as a new methodology to provide solution for existing problem and mainly operates on recurring problem to find reusable solution [12]. The particular design pattern once tested acts as a template for that specific problem henceforth.

2 Safety Quality Factors for Safety Critical Systems – An Overview

Certain factors are considered crucial in safety critical system that signifies safety of a system and such specific attributes should be consistently satisfied throughout the lifecycle of the system. Quality attributes characterize the various quality requirements expected in a system whereas safety attributes focuses more on the key factors that challenge safety of the system without which may lead to hazardous failures. These factors are manifested in this paper to enforce comparative analysis on the design pattern in the perspective of safety and to evaluate the extent of recommendation estimated compatible to safety critical system.

Safety is a kind of defensibility, dependability and also a kind of quality factor [15].Safety is a non functional requirement defined by MIL-STD-882D standard as Freedom from those conditions that can cause death, injury, occupational illness, damage to or loss of equipment or property, or damage to the environment [14]. Table 1 summarizes the safety quality factors reviewed in this paper to emphasize the support on safety factor for the design pattern used in real time safety critical systems. Among the factors mentioned PUF, RSI, SIL and reliability enumerate more significance to safety assessment whereas the other factors are less dependable as for as safety critical system is concerned.

Table 1. Safety quality factors for safety critical systems – an overview

Safety Factors	Description	Quantifiable Entity	Expected Value
Probability of Unsafe Failure (PUF)	PUF is calculated in relative to the probability of unsafe failure in a basic system that includes a single design channel and does not include any specific safety function [14].		Minimum
Relative Safety Improvement (RSI)	RSI defined as the percentage improvement in safety relative to the maximum possible improvement which can be achieved when the probability of unsafe failure is reduced to 0 [10].	$RSI = \left(1 - \frac{PUF(new)}{PUF(old)}\right) \times 100\%$ PUF(old) – PUF in basic system PUF(new) - PUF in the design pattern [10]	High
Safety Integrity Levels (SIL)	IEC61508 defines SIL as the probability of a safety-related system satisfactorily performing the required safety functions under all the stated conditions within a stated period of time [14].	SIL1,SIL2, SIL3 SIL4	SIL3 & SIL4
Mean Time to Failure (MTTF)	MTTF is the factor used with systems of high critical level in which the time between failures are calculated.		Low
Timing Performance	It is the temporal requirement that can be either the execution time or the deadline to be satisfied.		High
Availability	It is defined as the quantity of time that the system is functioning correctly and the proportion to which the system is available whenever needed.	$\alpha = \frac{MTTF}{MTTF + MTTR}$ MTTF-Mean Time to Failure MTTF-Mean Time to Repair [2]	High
Reliability	Reliability is defined as a characteristic of a component, expressed by the probability that the component will perform its required function under given conditions for stated time interval [2].		High
Modifiability	Modifiability of a software system is the estimation of the effect, cost and response of changes to the software.		High
Robustness	Robustness is the tolerance of the system to perform efficiently even in some faulty situation.		High

3 Design Pattern Supporting Safety Critical Systems

Memento, Façade, Proxy and Filter are the design patterns reflected in this context as these patterns support safety issues and requirements in addition to other application. The resultant framework in Fig.1 highlights the point of failure with the constructive measure executed at the failure adversity. It is inferred that though the four patterns are not designated exclusively for safety critical systems, supports the real time systems to some extent by satisfying certain attributes. Meanwhile, these patterns do not deliberate more on the key safety factors and hence it is acceded that the designated patterns are consented for systems with low critical level where only the basic quality attributes is required to be contented.

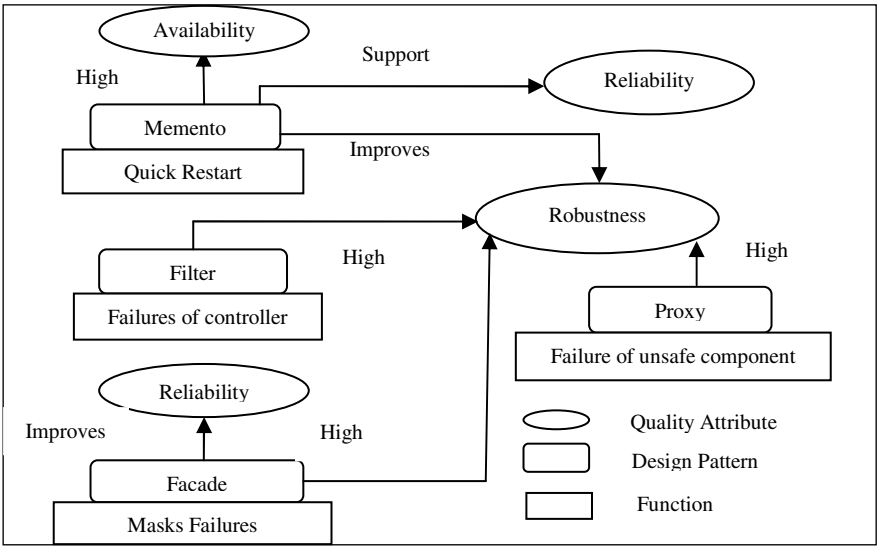


Fig. 1. Framework of Design Pattern based on quality attributes

3.1 N -Version Programming (NVP)

NVP is defined as the independent generation of $N \geq 2$ software modules, called versions, from the same initial specification [9]. In order to evaluate the support of safety parameters, the software safety simulator in which different versions are coerced to run in parallel on N hardware modules where voting technique is adapted to perform fault masking [10] is considered. To understand the correlation of quality attributes like availability and reliability the implementation result of NVP in multiple channel system [9] is deliberated.

Table 2 and Fig.2 framed by referring the analysis result reveals that, RSI value of NVP is condensed due to fault masking and is feasible to improve by introducing fault detection. PUF that is expected to be least is high in NVP due to the constraints that at least two versions to be correct to perform fault masking of a single fault. NVP imposes high reliability as it has provision of consuming any version as backup and

assures high availability as it operates on diversity principle. Since PUF is high, the RSI value is degraded and hence NVP is imperceptible to safety critical systems and incompetent to safety factors.

Table 2. Impact of Safety Factors on NVP

Safety Factors	Approximate % Supported by NVP
RSI	25
PUF	100
Availability	95
Reliability	90

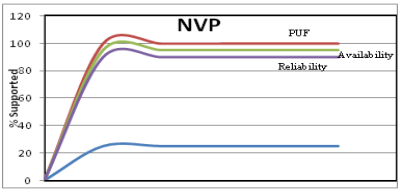


Fig. 2. Depiction of Safety Factors on NVP

3.2 Recovery Block (RB)

RB primarily concentrates on error detection and recovery mechanism and each recovery block comprises a primary block, an acceptance test, and zero or more alternate blocks [9].AT is executed towards the end of each version and if the test is successful it is considered as the final output, or else it is reversed back to original state and the similar procedure is repeated with another version. This process is accomplished until correct result is obtained or there are no more versions to be executed and in the worst case it is reported as overall system failure. The software safety simulator in [10] is abounded to assess the safety factors and to analyze the reliability, modeling the hard real time systems using fault analysis in [9] is considered where alternate blocks are used to identify the probability of failure. Substantiating to the analysis, Table 3 and Fig.3 exemplifies that the value of PUF is degraded, RSI is increased and when the failure rate increases the value of MTTF is decreased. Hence, RB can be used in safety critical system that may not adhere to the exact time bound and concerning reliability it can be designed and handled with a specific risk level that can be tolerable.

Table 3. Percentage of safety Factors supported by RB

Safety Factors	Approximate % Supported by RB
RSI	75
PUF	50
MTTF	25
Reliability	60

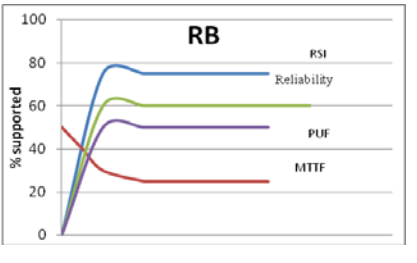


Fig. 3. Implication of Safety Factors on RB

3.3 Adaptive Control Filter (ACP)

ACP is used primarily for safety critical middleware systems and is designed to reduce the dependency of communication delay or packet loss on timing factor. Hence,

it primarily focuses on temporal requirements and more particularly on the command messages transmitted to the controller. The case study of Etherware in [8] is considered where the main function of ACP is to reduce timing dependency between application layer and communication layer. Whenever the delayed message due to network problem is discarded, ACP is devised to reduce the timing mismatch between components. The safety of ACP is achieved by assigning bounds or limit to the system thus making the system in safe state and hence the system should be coordinated within the safe zone and monitored.

Table 4. Quantity of safety supported by AVP

Safety Factors	Approximate % Supported by ACP
Reliability	50
Robustness	40
Timing Performance	100
Safety	60

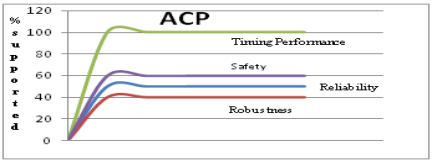


Fig. 4. Effect of Safety Factors on AVP

Since temporal requirements are highly satisfied, timing performance is high whereas the attributes reliability and robustness are average as depicted in Table 4 and Fig.4. Here overall safety is improved by making the system to work on safe state but also requires continuous monitoring and hence this pattern is more advantageous in area where control messages impact more and deadline of the message is assigned with high priority.

3.4 Acceptance Voting Pattern (AVP)

AVP is a hybrid pattern that incorporates the concept of NVP and RB Pattern [5].The key goal of AVP is to strengthen safety as well as reliability targeting the faults that remain even after software development. Emerging out with the primary version, the output of each version is forwarded to acceptance test to validate the output and if the result is valid, the same is forwarded as input to the dynamic voter. And this dynamic voter produces the correct output based on particular voting scheme and resolves the final output. The performance of non-functional requirements in [5] provides implications of reliability, safety and modifiability whereas the case study of software safety simulator [10] reveals the contribution of RSI and PUF.

Table 5 and Fig.5 represents RSI is highest in AVP since both fault detection and fault masking is embedded and PUF is reduced since, it is likely to generate correct result with single version. The reliability of AVP is magnified strictly based on the factors that the number of versions N should be consistent and the acceptance test essentially more effective. The safety integrity level highly recommended is SIL3 and SIL4, which is suggested ascertaining the diverse programming effect and fault detection accomplished using AT and dynamic voter. Since the result of each diverse version should be reviewed by AT, the waiting time of dynamic voter is relatively high which degrades the overall timing performance. In AVP the modification of single

version, AT and dynamic voter is reasonably uncomplicated but complex with N versions. Considering the overall response to safety quality factors AVP is suggested for systems that possess high level of safety criticality.

Table 5. Assessment of safety factors supported by AVP

Safety Factors	Approximate % Supported by AVP
RSI	100
PUF	50
SIL	SIL3 & SIL4
Timing Performance	60
Reliability	90
Modifiability	75

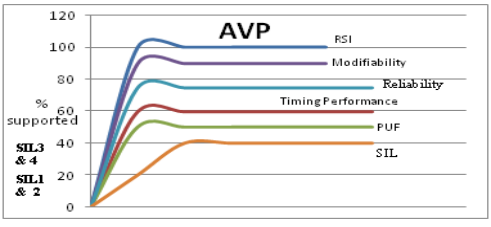


Fig. 5. Illustration of Safety Factors on AVP

3.5 Safety Executive Pattern (SEP)

SEP is generally consigned with complex systems where multiple safety concerns are addressed by a single system, fault detection is complex, or the fault recovery mechanisms are elaborate [11].The analysis of SEP in [5] is endured for evaluation where the context of the problem is the shutdown of prominent component might be unsuccessful or could be too long that may instigate to critical state. Hence, each time when the shutdown signal is persuaded this pattern checks whether it is in safe state and resolved by shutting down through the safety executive component if it is in safe state or else will switch over to other redundant units in case of failure.

Table 6. Quantity of safety supported by SEP

Safety Factors	Approximate % Supported by SEP
RSI	75
PUF	25
SIL	SIL3 and SIL4
Reliability	70
Modifiability	90
Timing Performance	100

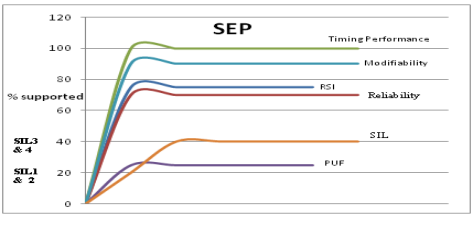


Fig. 6. Effect of Safety Factors on SEP

The result of case study in [11],[5] enclosed in Table 6 and Fig.6 symbolizes the quantifiable entities RSI,PUF and SIL .The improvement in RSI depends on the reliability of the safety executive and is highly recommended for SIL3 and SIL4.Since here all functions are executed in parallel due to the availability of resources, the timing performance is high and this pattern is recommended for composite systems in which complex fault recovery mechanism is required.

3.6 Recovery Block with Backup Voting (RBBV)

RBBV is a hybrid pattern that integrates the functionalities of NVP and RB pattern to construct efficient AT [13]. In RBBV, initially the basic version is executed followed by its acceptance test. If the initial version fails, a replica of the output is stored in the cache memory as a backup and the same process is continued with another version to execute the same functionality until any of the alternate versions passes the test or no more versions are available. The voter analysis the resultant values of the acceptance test and look for common values in the cache. If the common values are more it is considered as the final output of the acceptance test, or else it is decided as the problem with different versions and not with that of AT.

Table 7. Proportion of safety factors supported by RBBV

Safety Factors	Approximate % Supported by RBBV
RSI	70
PUF	50
SIL	SIL4 and SIL3
Reliability	90
Modifiability	100
Timing Performance	60

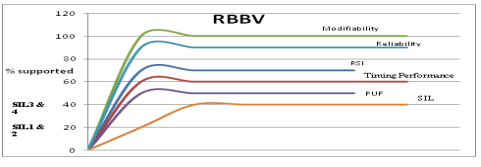


Fig. 7. Effect of safety factors on RBBV

For the analysis of safety factors, the case study of software safety simulator in [10] and representation of RBBV in [13] is considered. As depicted in Table 7 and Fig.7, RSI is average and PUF is and the reliability is .Based on the analysis done on RBBV for fault detection with different version, RBBV is highly recommended for integrity levels SIL4 and SIL3.

4 Analysis of Design Patterns for Safety Critical Systems

Table 8 represents the comparative analysis with respect to different safety quality metrics and in concern to safety critical systems, all the patterns favor safety factors to definite percent by supporting certain criteria to specific extent. It is observed that AVP and RRBV are the design patterns that satisfy all the safety factors in which AVP has magnified RSI, high reliability and distinguishable SIL level, suggested for high intensity safety critical systems. While concerning RBBV, the RSI value is above average with remarkable level of safety integrity. However, both the patterns satisfy all the safety criteria we imply RBBV as the suggestive pattern for safety critical system as it solves the problem of false negative cases by weak acceptance test by enchanting the quality of AT. Moreover, the reliability is high and all the safety concerns of NVP and RB is integrated in RBBV along with the additional safety measures.

The two factors to be reinforced in RBBV is RSI and timing performance in which the metric RSI is the core attribute for safety critical system and the degradation

happened due to the probability that may occur due to incorrect interpretation of the values in the cache. The common value in the cache may not be faulty always and in certain cases, it can be misinterpreted. This can be overwhelmed by concentrating more on the selection of version and rendering additional functionality to the dynamic voter to increase the value of RSI to the maximum. The timing performance of RBBV is adequate in the best case when the result is attained using single version. However, when it moves down to average case and worst case with different versions and voting technique, the timing performance of RBBV is less efficient. RBBV has overall average timing performance due to its sequential execution and execution overhead. The timing performance can be improved by upgrading with parallel execution and minimum version to reduce the execution overhead.

Table 8. Comparative Analysis of Design Pattern based on safety quality factors

Design Pattern / **Safety Factors**	**NVP**	**RB**	**ACP**	**AVP**	**SEP**	**RBBV**
RSI	Improves	Between NVP and AV		High		Between NVP and AV and Less than RB
PUF	High	Less than NVP		Less than NVP		Less than NVP
SIL				SIL3 & SIL4	SIL3 & SIL4	SIL3 & SIL4
MTTF		Less				
Timing Performance			High	Average	High	Average
Reliability	High	Average	Average	High	Greater than average	High
Availability	High					
Modifiability				Greater than average	High	High

5 Conclusion

Among the numerous design patterns intended for various applications, the suitable patterns for safety critical systems is exhibited here. Using relevant case studies the patterns are evaluated for the safety quality factors and the results are depicted. Six suitable design patterns are considered for analysis and the reviews are tabulated. It is observed that based on the contributing safety factors and quality attributes RBBV is the most suited design pattern for safety critical system. It adheres to all the safety criteria along with additional features whereas the safety factor RSI and temporal factor is to be improved to strengthen the safety level. Our future work will depend on the enhancement of RBBV for mission critical application and safety critical applications.

Acknowledgments. I am most grateful to my guide Dr.R.Selvarani, HOD (CSE), MSRIT, Bangalore, for the invaluable guidance, support and encouragement given throughout the preparation of this paper without which it may not have happened.

References

1. Shaw, M., Garlan, D.: Software Architecture: Perspectives on an Emerging Discipline. Prentice-Hall (1996)
2. Bass, L., Clements, P., Kazman, R.: Software Architecture in Practice. Addison-Wesley Professional (2003)
3. Fowler, Martin: Patterns of Enterprise Application Architecture. Addison-Wesley (2002)
4. Maier, M.W., Emery, D., Hilliard, R.: Software Architecture: Introducing IEEE Standard 1471. IEEE Computer 34(4), 107–109 (2001)
5. Armoush, A., Salewski, F., Kowalewski, S.: Design Pattern Representation for Safety-Critical Embedded Systems. J. Software Engineering & Applications (April 2009)
6. Wu, W., Kelly, T.: "Safety Tactics for Software Architecture Design. In: Proceedings of the 28th Annual International Computer Software and Applications Conference, COMPSAC 2004. IEEE (2004)
7. Mahemoff, M., Hussey, A., John, L.: Pattern-based Reuse of Successful Designs: Usability of Safety-Critical Systems. IEEE (2001)
8. Crenshaw, T.L., Robinson, C.L., Ding, H., Kumar, P.R., Sha, L.: A Pattern for Adaptive Behavior in Safety-Critical, Real-Time Middleware. In: Proceedings of the 27th IEEE International Real-Time Systems Symposium, RTSS 2006. IEEE (2006)
9. Avizienis, A.: The N-Version Approach to Fault-Tolerant Software. IEEE Transactions on Software Engineering SE-I 1(12) (December 1985)
10. Armoush, A., Beckschulze, E., Kowalewski, S.: " Safety Assessment of Design Patterns for Safety-Critical Embedded Systems. In: 35th Euromicro Conference on Software Engineering and Advanced Applications. IEEE (2009)
11. Douglass, B.P.: Real-Time Design Patterns. In: Real-Time UML:Developing Efficient Objects for Embedded Systems. Addison-Wesley (1998)
12. Sankar Ram, N., Rajalakshmi, B., Rodrigues, P.: Impact on Quality Attributes for Evaluating Software Architecture using ATAM and Design Patterns. Asian Journal of Information Technology 7, 126–129 (2008)
13. Armoush, A., Salewski, F., Kowalewski, S.: Recovery Block with Back up Voting:A New Pattern with Extended Representation for Safety Critical Embedded Systems. In: International Conference on Information Technology. IEEE (2008)
14. Armoush, A., Kowalewski, S.: Safety Recommendations for Safety-Critical Design Patterns. In: Proceedings of the International Workshop on the Design of Dependable Critical Systems (September 2009)
15. Firesmith, D.: Engineering Safety Requirements, Safety Constraints, and Safety-Critical Requirements. Journal of Object Technology 3(3), 27–42 (2004)

A Novel Scheme to Hide Identity Information in Mobile Captured Images

Anand Gupta[1], Ashita Dadlani[2], Rohit Malhotra[2], and Yatharth Bansal[2]

[1] Division of Computer Engineering
omaranand@nsitonline.in
[2] Division of Information Technology
{ashithadadlani,rohitmalhotra,yatharthbansal}@nsitonline.in

Abstract. A new research area concerning the application of steganography for hiding source identity information in mobile camera captured images is emerging. The work reported so far is focused on using MMS [Multimedia Messaging Service] and SMS [Short Message Service] as carrier media for covert communication and data security. The techniques employed for data hiding are lacking in a suitable combination of cryptography and steganography. The present paper proposes a novel scheme to hide source identity information (IMEI [International Mobile Equipment Identity] and IMSI [International Mobile Subscriber Identity] numbers) in mobile camera captured images. To achieve this, an application is developed that clicks images through mobile camera and hides IMEI and IMSI numbers in raw images before compressing them in PNG [Portable Network Graphics] format. The application encrypts the IMEI and IMSI numbers before hiding them. It uses a custom made key based algorithm for hiding these numbers randomly inside an image which ensures high security of hidden data. Runtime performance analysis of the above technique reveals that the computational lag due to encryption & steganography is miniscule. The technique is found feasible to run on actual mobile devices and can help identify the source of an anonymous mobile captured image.

1 Introduction

The process of capturing and sharing of data found, revolutionized with the advent of Smartphones. Smartphones contain camera and high speed Internet connectivity in the form of 3G and short range connectivity in the form of Bluetooth. This allows users to click images from the camera enabled Smartphone and quickly share them with their peers via bluetooth/internet etc.

The anonymity of source of such images can become a problem in certain judicial cases and in some common day to day scenarios too. In case someone receives an image from an anonymous e-mail id then the identity of the source of the image can't be determined. If the camera application itself embeds user information into the camera captured image then the source information gets attached to the image at the point of its production. This eliminates the need for complex trace-back in order to know the identity of the person whose mobile has been used to click that image. Imperceptibility is a key requirement. It means that the source information embedded

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 403–411.
springerlink.com

inside captured images must not be noticeable to anyone possessing that image. It is achieved using Steganography, which in Greek means 'covered writing'. Steganography is the art of hiding information such that it prevents the detection of hidden messages. It includes a vast array of secret communication methods that conceal the very existence of the messages. The methods include invisible inks, microdots, character arrangement, digital signatures, and covert channel and spread spectrum communications. The innocuous message used to hide the information is termed as 'carrier message'; the most commonly used carrier messages include images, audio and video. In this paper the carrier message is the image clicked by Smartphone camera and the message is user identity information in the form of IMEI-IMSI numbers.

IMEI stands for International Method Equipment Identity. It is a 15 digit code which is unique for every mobile set. IMSI stands for International Mobile Subscriber Identity. IMSI is a unique identification associated with all GSM [Global System for Mobile Communication] and UMTS [Universal Mobile Telecommunications Network] mobile phone users.

The research paper proposes a novel scheme to hide identity information in the form of IMEI and IMSI numbers inside images captured by the mobile phone camera. A sample application is designed and an algorithm to securely hide identity information in images is proposed. Performance of application in terms of time taken for hiding and security of hidden data is analyzed. The variation in time taken to hide with increasing length of data being embedded is also investigated.

Besides the present Section on Introduction, the paper is organized in the following manner. Related work done in the field of steganography in images is reviewed in Section 2 mentioning particularly the shortcomings in the approaches. Section 3 explains the motivation to propose an improved technique to overcome the identified shortcomings. Section 4 explains the proposed method in detail along with the system design. It includes a description of its implementation. Algorithms and the runtime analysis results are presented in Section 5 and Section 6 respectively to illustrate the dynamics and feasibility of the method. In the last Section 7, conclusion and scope for future work are discussed.

2 Related Work

Work done in the field of Steganography is either focused on Steganography techniques themselves or on their application to real life scenarios. Most of the research done earlier in this area proposes images, audios, and videos as cover media. Here the imperceptibility of hidden data is commonly achieved by exploiting the weaknesses of human auditory and visual systems, using the techniques for example, changing the least-significant bits of the pixels of a cover image to embed information.

The research paper in [1] discusses various steganography techniques for hiding data inside images. They cover least significant bit insertion and transform domain based steganography with special emphasis on the comparison between RSA and elliptic curve based digital signatures.

The paper in [2] develops over the commonly used LSB [Least Significant Bit] replacement technique by randomizing the distribution of data by first dividing the

image into small blocks and then selecting pixels from the block based on a password for LSB [Least Significant Bit] insertion. The method requires the image to be loaded in a block wise manner and requires extra memory space for keeping note of the selected pixels in a block.

In [3], the technique discussed in [2] is applied for hiding information in MMS [Multimedia Messaging Service].

The carrier medium in [3] consists of text and image, steganography is applied in these components to accomplish the purpose.

3 Motivation and Contribution

The motivation behind this work is the possibility to apply Steganography to Smartphone images. The utility of the same is discussed in Section I of this paper. The technique discussed here improves upon the method developed in [2] as –

1. It embeds one bit per pixel. The capacity for hiding (the number of bits of message data that can be hidden in a carrier image) is reduced in comparison to that obtained by the method developed in [2], because a full pixel is now utilized for hiding just one bit of message information. But the changes in carrier image are less noticeable, because the spread of message information is now wider.
2. It encrypts the message with a 128 bit key using AES [Advanced Encryption Standard] prior to embedding it inside the camera captured image.
3. It uses a key whose length is equal to that of the data being embedded which in the particular case of hiding IMEI (15 bytes) and IMSI (15 bytes) numbers is 248 bits. This key is used to randomly distribute data in the image and recover them afterwards.

A new algorithm is introduced which hops over pixels in the carrier image based on a hop sequence depending on the 248 bit key described above. As it can be seen, the keys and algorithms used for encryption and random distribution are different. Considering the length of encryption and embedding keys used, there are a total of $2^{128} * 2^{248}$ key combinations possible which make it nearly impossible to crack the message using brute force or dictionary attack.

Also since the routine responsible for the embedding process is part of the camera application itself, the operation underway goes unnoticed by the user.

4 System Design

The platform of choice is Android, known for its open software stack and wide reception by the public.

Eclipse combined with the Android Development Toolkit (ADT) and Android API provides a convenient interface for developing android applications.

In this section the architecture of the software system is discussed. The architecture is explained below (See Fig.1):–

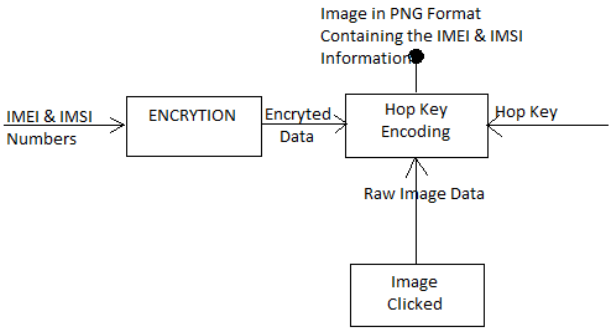


Fig. 1. Schematic for the proposed approach

The system consists of two main components –

1. **The Camera Application** – The job of this application is twofold –
   - To interface with the camera to take images
   - To hide user identity information inside the images after encryption and save the image after compression. PNG format is used as it's lossless, so the embedded information is not lost after compression.
2. **The Decoder Application** - This application works independently from the camera application and its task is to take compressed PNG images as input and extract-decrypt the user identity information numbers using the same keys as used by the encoder.

Both the above applications are developed for the android platform using Java.

The Camera Application mainly consists of (See Fig.2):–

- **Image Capturing Code** – This is the portion responsible for interfacing with the camera and setting up surfaces for image capture. It sets up callbacks to 'click' events, initializes and maintains the camera object and fetches raw image data once an image has been captured. It also fetches the IMEI-IMSI numbers using the telephony manager object. After steganography has been performed on the raw pixel array, it performs compression and saves it in the PNG format.
- **AES Encryption** - This portion takes as its input the IMEI-IMSI numbers and performs AES encryption using a 128 bit key. The output is encrypted IMEI-IMSI numbers.
- **Encoders** – These portions takes as its inputs, the raw pixel array, hop key and encrypted IMEI-IMSI numbers and returns a pixel array in which the encrypted numbers are hidden.

The sequences followed are,

Encryption (encrypting IMEI-IMSI numbers) -> Encoding (embedding encrypted IMEI-IMSI numbers in image) -> Storing Image.

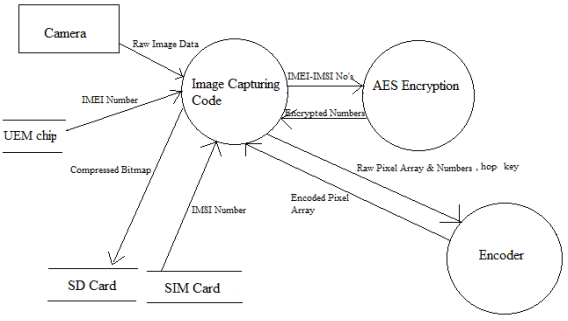


Fig. 2. Data flow diagram of the camera application (encoder)

Similarly the Decoder application consists of the following components (See Fig.3):–

- **Image Fetching Code** – It is responsible for fetching an image from the SD card and sending it over to the decryption and decoding routines. Upon receiving the decoded IMEI-IMSI numbers it prints them on the screen.
- **AES Decryption** – It performs decryption on the encrypted numbers that are fetched from the image.
- **Decoder** – The decoder takes the pixel array and hop key as its inputs and delivers the encrypted IMEI-IMSI numbers as its output.

In the decoder application the steps undertaken are in a sequence opposite to that in the camera application (encoder).

Image Fetching -> Decoding (retrieval of encrypted IMEI-IMSI numbers) -> Decryption. The keys used for decryption and hopping are exactly the same as those used by the camera application (Symmetric keys).

The data flow diagram of the decoder application is given in Fig.3. It illustrates the flow of data between various components of the application.

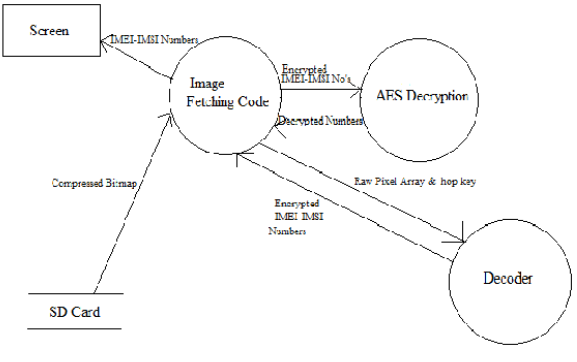


Fig. 3. Data flow diagram of the decoder application

5 The Algorithm

The Algorithms used in this paper are –

1. **AES** – It is a standard symmetric key encryption algorithm, a block cipher. The implementation used is the one provided in the java cryptography package, javax.crypto.
2. **Hop Key Algorithm** – It is the novel algorithm developed in this paper for distributing data inside the image. The key used by this algorithm is the 'hop key', a key whose length is same as that of the Data, which in this case is 248 bits. An offset variable is kept which points to the pixel in which the present bit of message needs to be hidden. The algorithm is described below –

ALGORITHM – *Hop Key Algorithm ()*

1. Consider each byte of the message that needs to be hidden in the pixel array.
2. Consider a byte of the hop key alongside what was done in step 1.
3. Progress bitwise in both the bytes, if the current bit in key is 0 then hide message bit in the position indicated by offset variable; else if current bit is 1 then hide the message in the position indicated by offset variable + 1.
4. Increment the offset variable by 1.
5. Repeat steps 3 to 4 for each bit in message byte.
6. Repeat steps 1 to 5 for each byte in message array.

The same algorithm is used by the decoder application to extract the encrypted numbers from within the image.

6 Performance Analysis and Result

The code is run on android emulator under eclipse. The embedding of IMEI and IMSI numbers takes place just after the user clicks his mouse/touchpad, after which the image gets stored in secondary storage.

AES v/s DES

Two symmetric key encryption algorithms are being considered as alternatives here- Data Encryption Standard and Advanced Encryption Standard.

Key length of AES is however much more than that of DES. The former has key lengths of 128/192/256 bits available with it while the latter only has 56 & 64 bit ones.

A comparison of the time taken by DES and AES to encrypt the same length of data is shown in Table –1.

Table 1. Comparison of Time Performance between AES & DES

DES (in ns)	392423	415288	434418	413888	369559
AES (in ns)	315432	279036	324298	270637	280436

The median time of performance of DES is 413888 nanoseconds.
The median time of performance of AES is 279036 nanoseconds.
It's clear that AES is not only more secure because of its bigger key length but is also faster at encryption than DES.

Key Based Encoding v/s Sequential Encoding

It is also tested to see whether the key based hop distribution of data within the image imposes any significant additional time overhead over simple sequential distribution of data within the image.

Table 2. Time Performance (Key v/s Sequential Encoding)

KEY BASED (ns)	357427	430219	432552	384957	396156
SEQUENTIAL (ns)	344772	416164	312575	308843	357370

Median time for Key based encoding – 396156 ns
Median time for Sequential based encoding – 322841 ns
Difference – 72315 ns = 0.07 milli seconds

The difference is miniscule and will go unnoticed by the human eye. The time lag is insignificant. The sequential system of embedding data embeds message data in a continuous region inside the carrier image, which makes it easier to get the embedded message as one just has to check for every offset possible in the image. The embedded message consists of IMEI and IMSI numbers in an encrypted form. Since these numbers are available to anyone possessing the mobile, the embedded message can be subjected to 'known plain text' attack. The key based system adds an additional level of security over encryption by distributing the encrypted IMEI and IMSI numbers in the carrier image based on a key. The key's length is same as that of the message data which in this case is 248 bits long. Thus in order to get encrypted IMEI and IMSI numbers an attacker will have to breach this 248 bit long secret key. It makes even harder to get the encrypted data. 1 in $2^{248} * 2^{128}$ are the odds of an attacker correctly getting the IMEI and IMSI numbers hidden in the image.

Table below shows variation in time taken to perform encoding against the change in length of data being embedded.

Table 3. Variation in Time Performance with Length of Data

LENGTH(bytes)	1	2	4	6	8
TIME (ns)	105455	111988	155849	165648	170781

As shown in table 3 the time taken to encode increases steadily with increase in size of data being embedded. The size of jumps taken at each step shows no discernible pattern, but a monotonic increase is observed.

7 Comments on Results

The results presented in the previous article can be summed up as follows –

I. The time performance of AES is far superior compared to DES. Also the security provided by AES is better because of longer key length. AES has 128/192/256 bit keys while DES offers only 56 & 64 bit keys.

II. The statistical difference between time taken for key based encoding and sequential encoding (straight embedding without using the hop key) is very less. This suggests that key based encoding is embedding the data into image without imposing significant performance overhead.

III. The time taken to embed data into an image increases monotonically with increase in size of data being embedded.

IV. Encryption with 128 bit AES key prior to encoding ensures a 1 in 2^128 chance for retrieval of data. Further spread into the image pixels using a 248 bit key ensures a 1 in 2^248 bit probability of finding the embedded encrypted data. The combined probability of precisely discovering hidden data is brought down to 1 in 2^128 * 2^248.

8 Conclusions and Future Scope

The paper presented a technique to hide identity information inside mobile captured images. Technique proposed key based distribution of data inside image. Key based distribution promises significant benefits over sequential distribution in terms of security. Time performance of both methods is comparable.

Steganography in smartphones is a relatively new idea and the work done in the present paper can be further extended by–

1. Deployment of the application on android based mobile devices. Taking measures to ensure that the IMEI and IMSI numbers embedded in an encrypted form are robust to minor bit level corruptions in the carrier image.
2. Further randomizing the distribution of data in image by coupling hop size with random number generator. It will improve the security greatly.
3. Testing the performance of encoding algorithm with the change in image size.

The scope of steganography in smartphones is huge and the idea of hiding user identity with data produced by his/her phone can have wide implications to several cases of dispute.

References

[1] Lenti, J.: Steganographic Methods. Periodica Polytechnica Ser. El. Eng. 44(3-4), 249–258 (2000)

[2] Shirali Shahreza, M.: An Improved Method for Steganography on Mobile Phone. In: Sandberg, I.W. (ed.) Proceedings of the 9th WSEAS International Conference on Systems, ICS 2005, World Scientific and Engineering Academy and Society (WSEAS), Article 28, 3 pages, Stevens Point, Wisconsin, USA (2005)

[3] Shirali Shahreza, M.: Steganography in MMS. In: Proceedings of the 11th IEEE International Multitopic Conference, Lahore, Pakistan, December 28-30 (2007)
[4] Dhanashri, D., Patil Babaso, S., Patil Shubhangi, H.: Mms Steganography For Smartphone Devices. In: Proceeding of 2nd International Conference on Computer Engineering and Technology, Jodhpur, India, November 13-14, vol. 4, pp. V4-513–V4-516 (2010)
[5] Gupta, A., Barr, D.K., Sharma, D.: Mitigating the Degenerations in Microsoft Word, Documents: An Improved Steganographic Method. In: Proceedings of the 2nd International Conference on Computer, Control and Communication (IC4 2009), Karachi, Pakistan, February 17-18, pp. 1–6 (2009)
[6] Dobsicek, M.: Modern Steganography, Department of Computer Science and Engineering, Faculty of Electrical Engineering, Czech Technical University, Prague (2003), http://www.scycore.com/papers/ow04_paper.pdf

A License Plate Detection Algorithm Using Edge Features

Mostafa Ayoubi Mobarhan, Asadollah Shahbahrami, Saman Parva,
Mina Naghash Asadi, and Atefeh Ahmadnya Khajekini

Computer Engineering Department, Faculty of Engineering, University of Guilan, Rasht, Iran
shahbahrami@guilan.ac.ir, mostafa.ayoubi7@yahoo.com
saman.parva@gmail.com, Asadi.mina@gmail.com,
atefeh.a777@yahoo.com

Abstract. License Plate Detection (LPD) is a main step in an intelligent traffic management system. Based on many techniques have been proposed to extract license plate of different vehicles in different condition. Finding a technique that provides a good accuracy with a good time response time is difficult. In this paper, we propose a technique for LPD. The proposed technique uses the edge features to find the rows position of the license plate. In order to find the column positions, we find the small blue color part at the left of the license plates in the determined rows position. Our experimental results for different image database show that an accuracy of 96.6 percent can be achieved.

Keywords: License Plate Detection, Edge Detection.

1 Introduction

Intelligent traffic management system controls different vehicles on roads with new technology of computers and communication systems. This traffic control system has many parts such as controlling traffic lights and automatic vehicle identification [1]. Automatic License Plate Identification (LPI) is an essential stage in the automatic vehicle identification.

Generally, LPI has three major parts, License Plate Detection (LPD), character segmentation, and character recognition [22,23]. In the first step, the position of the plate is determined in an image. In order to recognize different characters, character segmentation algorithms are applied. The accuracy of the second and third stages depends on the first stage, extraction of plate. In other words, the main difficult task in LPI is the LPD. This is because of the following reasons. First, license plates normally, occupy a small portion of the whole image. Second, license plates have different formats, styles, and colors.

One way to implement the LPI is using digital image processing technology [2, 3]. In this paper, we propose a technique for LPD in an image. First, the algorithm finds the rows position of the license plates by using vertical edge features. Second, in order to find the column positions of the license plates, the proposed algorithm finds small blue color section at the left side of the license plates in the previously determined rows positions. This improves the performance because it is not necessary to search whole RGB image to find the blue part.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 413–421.
springerlink.com

This paper is organized as follows. In Section 2 we present a briefly explanation of LPI. The proposed technique is discussed in Section 3 followed by experimental results in Section 4. Finally, paper ends with some conclusions in Section 5.

2 License Plate Identification

There are different stages for LPI [20, 21, 22, 23, 24, 25]. For example, in [22], the LPI system has been divided into three stages, LPD, character segmentation and character recognition.

A desired LPI system has to work under different imaging conditions such as low contrast, blurring and noisy images. This is because images that contain license plates are normally collected in different conditions such as day, night, rainy, and sunny.

Different researchers used different tools for this purpose. For example, neural networks tool has been used in [4], and pattern matching [5], edge analysis [6], color and fuzzy maps [7], [8], vector quantization, texture analysis [9], Hough transform [21][10], dynamic programming-based algorithms [11], corner template matching[12], morphologic techniques [13,14], IFT-based fast method for extracting the license plate[15] have been applied. In addition, different platforms are used to implement LPI [16, 17, 18, 19]. In all these algorithms finding a technique that provides a good accuracy with an acceptable response time is difficult.

3 The Proposed Algorithm

We propose a technique for LPD in this section. The proposed algorithm has different stages, color space conversion, edge detection and image binarization, finding appropriate rows which contain license plate, finding blue region in original image.

Input images are in RGB format. In order to reduce image size and increase the processing time, we convert the RGB images to gray scale images by using Equation 1 [26]:

$$Y = 0.299 * R + 0.587 * G + 0.114 * B \tag{1}$$

Some of these RGB images and their corresponding gray scale images are depicted in Figure 1.

The plate regions in images usually have more edges than other area. In other words, these areas are more crowded than other area. This is because we have different numbers, characters, and colors close to each other in plate regions. We use this feature to extract the row position of plates. In order to this purpose we implement an edge detector such as Sobel to obtain the gradient of the input image in vertical direction. The mask of the Sobel edge detector is depicted in Equation 2 [27, 28, 29].

$$h = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \tag{2}$$

Fig. 1. Conversion of the RGB images to gray scale images

After applying edge detection algorithm, we convert output image to a binary image. Those pixels which belong to edges are assigned to value one and other pixels are assigned to zero. Figure 2 depicts some outputs of this stage.

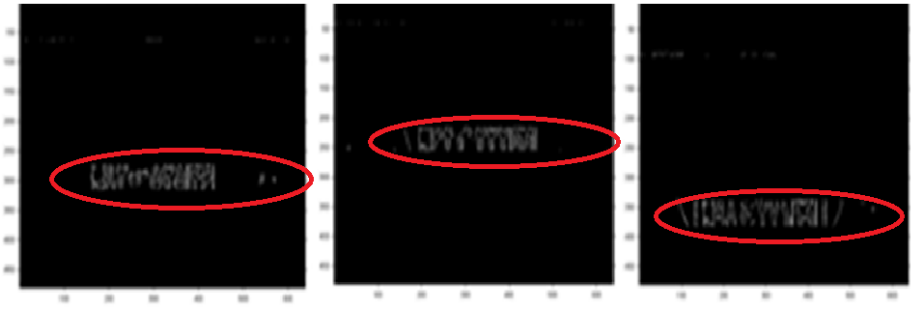


Fig. 2. Appling the edge detection algorithm and image binarization

In the third step of the proposed technique, we find the row position of plats by using the binary images. For this purpose, we count the number of ones, zeros, and their changes; one to zero and zero to one for each row. A predefined threshold has been set for this counted numbers. If the counted numbers for a row are greater than the threshold then that row belongs to a plate region, or our plate region is located in that row. The implemented algorithm is depicted in Figure 3.

```
// Extraction of candidate rows
For each row of image matrix {
   Calculate sum of pixel values
   Calculate changes of neighbour pixels
}
For each rows of image matrix {
   If sum of this row <a threshold value & change
      number of neighbour pixels <another threshold
      value of whole pixel that belong to this row=0
   Else
      This row is nominated for plate area
}
Start row of candidate region =first row that has the previous step
conditions
While isn't recognized the candidate region {
    While distance between next row and start row<a threshold
       end row=next row
a threshold >    If distance between end row and start row
      Break from loop
   Else {
      Start row=end row
      Continue the loop
}
```

Fig. 3. Proposed code for extraction of candidate rows

Finding the plate position in a binary, edge image is so fast with the mentioned algorithm. After this step, we should find the column position of the plate region. In order to perform this stage, we find the blue color of the plate in the original image. This is because all Iranian plates have a small blue color part at the beginning of the plates. We use previous results, the position of rows in the binary image for corresponding rows in the original image. In other words, in order to find the blue color part of the plates in whole RGB images, we find it in a small area of the RGB images. This is because we already know the rows position and this idea improves the performance of the proposed algorithm. The implemented algorithm for this stage is depicted in Figure 4.

```
// Extraction of blue regions
While isn't recognized the candidate blue area of plate {
  For each row of RGB image that belong to candidate area {
    For each column of image {
      If result of subtraction value of pixel from desired value<a threshold
        For next pixel until n pixel after it {
          If result of subtraction value of pixel from desired value<a threshold
            Number of continuous blue pixel of this row++
          Else
            Break from this ring
        }
      If number of neighbour pixels that have blue value>a threshold {
        This row belong to blue box of plate area
        If number of neighbour blue pixels of this row> number of neighbour blue pixels of previous rows
          Width of blue box of plate = number of neighbour blue pixels of this row
        Break from column loop and go to next row
      }
      Else
        Continue to examine RGB values of next pixels of present row
    }
  }
  If height of present area<a threshold value
  Repeat the previous steps for recognition a blue region but column loop start from end column of present area
Else {
  Execute horizontal Sobel inside the present area
  Execute vertical Sobel inside the neighbour region with the same size of present area
  Execute horizontal Sobel inside the present area
  If result of these more than specified thresholds
    This area is the same blue box of plate
  Else
    Repeat the previous steps for recognition a blue region but column loop start from end column of present area
}
}
Detection of plate region from blue area situation
```

Fig. 4. Proposed code for extraction of blue regions

After finding the small blue part of the plate region, we can find the whole part of plate using the ratio of wide to height of the plate. All input image samples have been taken in distance of one and half meters in the rear of cars. The ratio of wide to height of the plats is set as a threshold. The results of this stage are depicted in Figure 5.

Fig. 5. Plate extraction

4 Evaluation Results

We use 150 car images that contain different license plates. The image resolution is 640 X 480 pixels. We captured them using A4 TECH USB2.0 PC Camera. Our images acquired from about 1.5 meters away from the rear of the vehicle. To improve the complexity and universality of the test databases, the images are acquired from a large traffic crossing with different lightening conditions, sunny, cloudy, shadow, daytime, and nighttime, for different kinds of Iranian vehicle such as van, truck, and car.

The implementation results are depicts in Table 1. As this table shows the number of correct detection in images that have been taken in day and night time is 141 and 145 out of 150, respectively. In addition, percentage of accuracy is also 94 and 96, respectively.

Table 1. Evaluation results for different images that have been taken in day and night time

Lightening conditions	Number of correct detection	Percentage of Accuracy (%)
LPD Performance in day	141/150	94%
LPD Performance in night	145/150	96.6%

Figures 6 and 7 depict some sample outputs which have been obtained using the proposed algorithm for both images, day and night time.

In addition, we measured the execution time of the proposed technique. Execution time for both image sets, day and night images is 202 Millisecond.

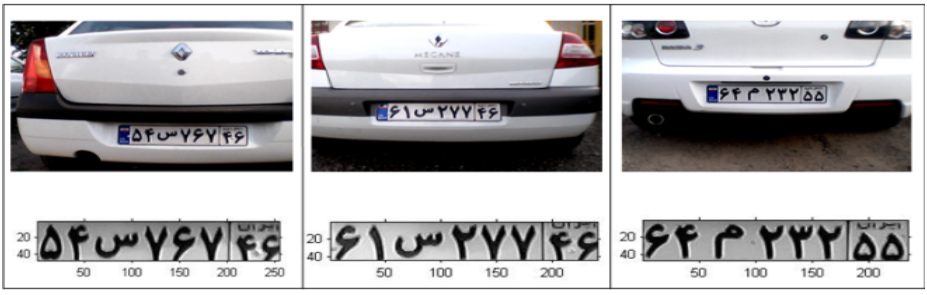


Fig. 6. Examples of extraction license plate using the proposed algorithm in images which have been taken in day time

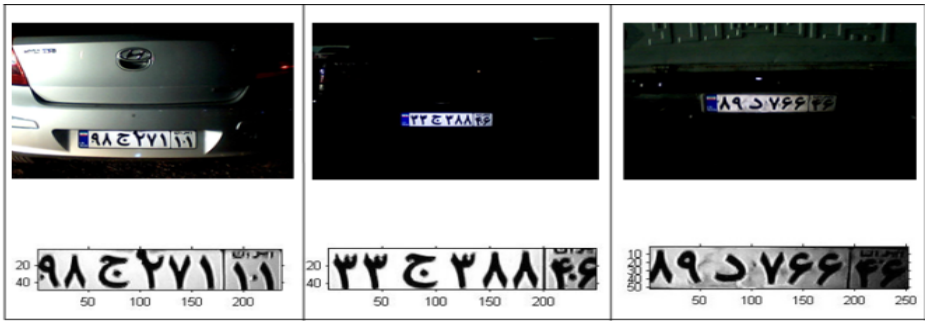


Fig. 7. Examples of extraction license plate using the proposed algorithm in images which have been taken in night time

5 Conclusions

In this paper, we presented a technique to detect license plates in images which have been taken in different conditions from Iranian vehicles. The proposed algorithm has many stages, color space conversion, applying an edge detection algorithm and image binarization, finding rows and columns position of the plates. The algorithm was tested over a large number of images. The accuracy of the proposed algorithm is 94% and 96.6% for images which have been taken in day and night, respectively. Our proposed algorithm needs just 202 ms to extract plates region.

References

[1] Ozbay, S., Ercelebi, E.: Automatic Vehicle Identification by Plate Recognition. In: World Academy of Science Engineering and Technology, pp. 222–225 (2005)

[2] Wu, M.K., Wei, J.S., Shih, H.C.: License Plate Detection Based on 2-Level 2D Haar Wavelet Transform and Edge Density Verification. In: IEEE Int. Symposium on Industrial Electronics, pp. 1699–1704 (2009)

[3] Nguyen, C., Ardabilian, M., Chen, L.: Unifying Approach for Fast License Plate Localization and Super-Resolution. In: Int. Conf. on Pattern Recognition, pp. 376–379 (2010)
[4] Akoum, A., Daya, B., Chauvet, P.: Two Neural Networks for License Number Plates Recognition. Journal of Theoretical and Applied Information Technology, 25–32 (2005)
[5] Irecki, D., Bailey, D.G.: Vehicle Registration Plate Localization and Recognition. In: Proc. of the Electronics New Zealand Conf. ENZCon'OI, New Plymouth, New Zealand, pp. 236–247 (2001)
[6] Deb, K., Chae, H., Jo, K.: Vehicle License Plate Detection Method Based on Sliding Concentric Windows and Histogram. Journal of Computers 4(8), 771–777 (2009)
[7] Chang, S., Chen, L., Chung, Y., Chen, S.: Automatic License Plate Recognition. In: IEEE Int. Conf. Intelligent Transportation Systems, pp. 236–249. IEEE Computer Society Press (2004)
[8] Shi, X., Zhao, W., Shen, Y.: Automatic License Plate Recognition System Based on Color Image Processing. In: Gervasi, O., Gavrilova, M.L., Kumar, V., Laganá, A., Lee, H.P., Mun, Y., Taniar, D., Tan, C.J.K. (eds.) ICCSA 2005, Part IV. LNCS, vol. 3483, pp. 1159–1168. Springer, Heidelberg (2005)
[9] Cano, J., Perez-cortes, J.C.: Vehicle License Plate Segmentation in Natural Images. In: Perales, F.J., Campilho, A.C., Pérez, N., Sanfeliu, A. (eds.) IbPRIA 2003. LNCS, vol. 2652, pp. 142–149. Springer, Heidelberg (2003)
[10] Yanamura, Y., Goto, M., Nishiyama, D., Soga, M., Nakatani, H., Saji, H.: Extraction and Tracking of The License Plate Using Hough Transform and Voted Block Matching. In: IEEE Intelligent Vehicles Symposium, pp. 243–246 (2003)
[11] Kang, D.: Dynamic Programming Based Method for Extraction of License Plate Numbers of Speeding Vehicles on The Highway. Journal of Automotive Technology 10(2), 205–210 (2009)
[12] Tahir, A., Habib, H.A., Khan, M.F.: License Plate Recognition Algorithm for Pakistani License Plates. Canadian Journal on Image Processing and Computer Vision 1(2), 16–21 (2010)
[13] Zamani, A., Movahedi, M.: Car Plate Identification Using Morphology and Constant Moments Transform. In: Conf. Ferdowsi University, Mashhad, Iran (2007)
[14] Kasaei, H., Kasaei, M., Kasaei, A.: New Morphology-Based Method for Robust Iranian Car Plate Detection and Recognition. Journal of Computer Theory and Engineering 2(2), 264–268 (2010)
[15] Abolghasemi, M., Ahmadifard, A.: A Car Plate Identification System Base on LFT Transform. In: Conf. Ferdowsi University, Mashhad, Iran (2007)
[16] Kamat, V., Ganesan, S.: An Efficient Implementation of the Hough Transform for Detecting Vehicle License Plates Using DSP's. In: IEEE Real-Time Technology and Application Symposium, RTAS, pp. 58–72 (1995)
[17] Kang, J., Kang, M., Park, C., Kim, J., Choi, Y.: Implementation of Embedded System for Vehicle Tracking and License Plates Recognition Using Spatial Relative Distance. In: 26th Int. Conf. on Information Technology Interfaces, pp. 167–172 (2004)
[18] Bellas, N., Chai, S.M., Dwyer, M., Linzmeier, D.: FPGA Implementation of A License Plate Recognition Soc Using Automatically Generated Streaming Accelerators. In: 20th Int. Parallel and Distributed Processing Symposium, IPDPS, pp. 8–19 (2006)
[19] Park, E.A.: OCR in A Hierarchical Feature Space. IEEE Transactions on Pattern Analysis and Machine Intelligence, 400–407 (2000)

[20] Bailey, D.G., Irecki, D., Lim, B.K., Yang, L.: Test Bed for Number Plate Recognition Applications. In: Proc. of the First IEEE International Workshop on Electronic Design, Test and Applications, DELTA 2002, pp. 501–503 (2002)
[21] Chen, Z., Wang, G., Liu, J., Liu, C.: Automatic License Plate Location and Recognition Based on Feature Salience. Int. Journal of Computational Cognition 5(2), 1–9 (2007)
[22] Mahini, H., Kasaei, S., Dorri, F.: An Efficient Features-Based License Plate Localization Method. In: Proc. of IEEE Int. Conf on Pattern Recognition, ICPR, pp. 841–844 (2006)
[23] Zheng, D., Zhao, Y., Wang, J.: An Efficient Method of License Plate Location. Pattern Recognition Letters, 2431–2438 (2005)
[24] Pan, X., Ye, X., Zhang, S.: A Hybrid Method for Robust Car Plate Character Recognition. Engineering Applications of Artificial Intelligence, 963–972 (2005)
[25] Parasuraman, K., Kumar, P.: An Efficient Method for Indian Vehicle License Plate Extraction and Character Segmentation. In: IEEE Int. Conf. on Machine Vision and Humanmachine Interface, pp. 447–452 (2010)
[26] Duan, J., Qiu, G.: Novel Histogram Processing for Color Image Enhancement. In: IEEE Int. Conf. Image Graph, pp. 55–58 (2004)
[27] Abolghasemi, V., Ahmadyfard, A.: A Fast Algorithm for License Plate Detection. In: Conf. on Advances in Visual Information Systems, pp. 468–477 (2007)
[28] Abolghasemi, V., Ahmadyfard, A.: An Edge-based Color-aided Method for License Plate Detection. Image and Vision Computing, 1134–1142 (2009)
[29] Kolour, H.S., Shahbahrami, A.: An Evaluation of License Plate Recognition Algorithms. International Journal of Digital Information and Wireless Communications (IJDIWC), 247–253 (2011)

Human and Automatic Evaluation of English to Hindi Machine Translation Systems

Nisheeth Joshi[1], Hemant Darbari[2], and Iti Mathur[1]

[1] Apaji Institute, Banasthali University, Rajasthan, India
[2] Centre for Development of Advanced Computing, Pune, Maharashtra, India
nisheeth.joshi@rediffmail.com, darbari@cdac.in,
mathur_iti@rediffmail.com

Abstract. Machine Translation Evaluation is the most formidable activity in Machine Translation Development. We present the MT evaluation results of some of the machine translators available online for English-Hindi machine translation. The systems are measured on automatic evaluation metrics and human subjectivity measures.

Keywords: Machine Translation Evaluation, Subjective Evaluation, BLEU, METEOR.

1 Introduction

Ever since the research in the field of machine translation (MT) has started, evaluation of machine translation (MT) engines has been the most formidable activity. Miller & Beebe-Center in 1956 and Pfafflin in 1965 were the first researchers who proposed the strategies to evaluate MT Systems. This was the time when all the evaluation was dependent on human evaluator's judgements. This approach was fairly effective, but was very time consuming, as human evaluators took days, sometimes months to evaluate hundreds or thousands of sentences generated by MTs. With time this approach was improved by several researchers.

Automatic evaluation of MT engines began with the introduction of BLEU Metric which was developed by Papineni et al (2001). They proposed measures to automatically evaluate MT Output based on n-gram approach. Since then there have been a lot of research in this new paradigm. Today, we have a plethora of evaluation metrics based on this approach which can be employed to evaluate MT Systems.

In this paper we discuss the amalgamation of human and automatic evaluation methods. We have conducted our study on some of the MT engines available for English-Hindi language pair. In section 2, we briefly discuss human and automatic evaluation measures. In this section we have also provided a brief review of the work done in the area of human and automatic evaluation of MT systems. Section 3 describes our evaluation methodology. Section 4 shows the result and analysis of our study and finally section 5 concludes the work done.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 423–432.
springerlink.com

2 Approaches to Machine Translation Evaluation

Broadly, the entire MT Evaluation arena can be divided into two categories, Human Evaluation and Automatic Evaluation. They are discussed in the following section.

2.1 Human Evaluation

Human or manual evaluation is considered as the golden metric of MT evaluation. It allows MT developers to measure the quality of their system on a wide range of parameters. Many approaches of human evaluation have been described in literature. Lehrberger & Bourbeau (1988) described a general methodology for performing evaluation of MT systems. Dabbadie et al (2002) described a detailed approach for performing reliability tests. Falkedal (1994), and Arnold et al (1994) summarized various human evaluation measures that were developed and used by the evaluators of their times. On a more recent account, Wilks (2008) conducted a study in which he show that a human evaluator who only has knowledge of the target language can produce similar evaluation scores as compared to the one who has knowledge of both source and target languages. Callison-Burch et al (2007) conducted informativeness evaluation of MT engine outputs in which they asked the evaluators to edit the translations, before letting them see the reference translations.

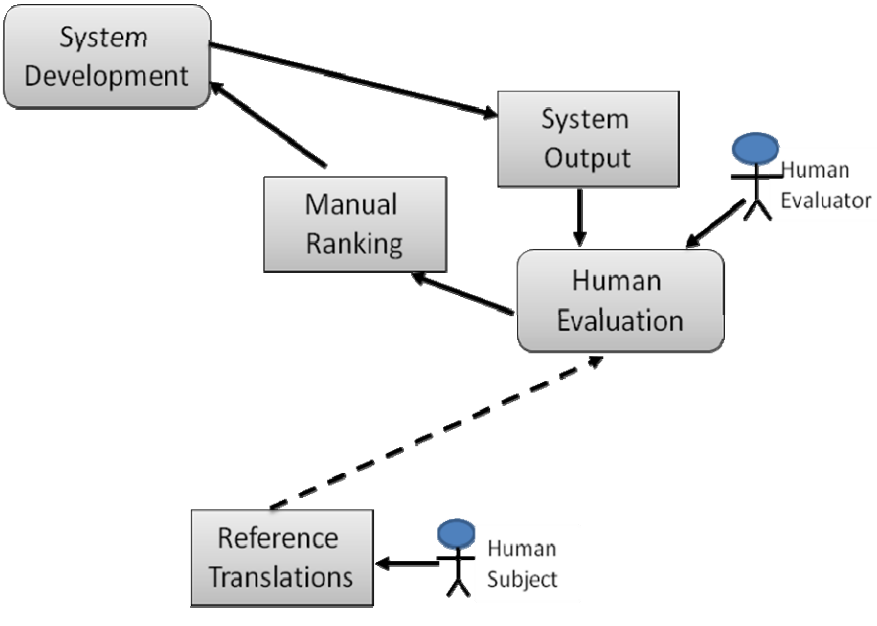


Fig. 1. Human Evaluation Process

In human evaluation, human evaluators look at the output and judge them by hand to check whether it is correct or not. Bilingual human evaluators who can understand

both input and output are the best qualified judges for the task. Figure 1 shows the process of human evaluation.

Here, the output of the system is provided to the human evaluator, who evaluates the output on the basis of a subjective questionnaire/metric, based on which an evaluator can judge the output. This is done for each of the sentences which are going to be evaluated. But, judging MT output merely on the basis of correctness is a very harsh method of evaluation. So, MT Researchers developed metrics onto which an output can be checked for fluency (grammatical and idiomatic word choice correctness) and adequacy (meaning preservation).

These two factors can be incorporated in a single metric or can be provided as a separate metrics. Moreover a human evaluator has to adjudge the output on the basis of some scale which may range from metric to metric. For example Figure 2 shows adequacy scale for evaluation of sentences and Figure 3 shows fluency scale.

Adequacy	
4	Completely Meaningful
3	Partially Meaningful
2	Little Meaningful
1	None

Fig. 2. Four Point Scale for Adequacy

Fluency	
4	Completely Comprehensible
3	Partially Comprehensible
2	Little Comprehensible
1	Incomprehensible

Fig. 3. Four Point Scale for Fluency

In spite of their informative capabilities, human evaluation metrics have several limitations. Some of them are:

2.1.1 Slow and Expensive

Human evaluation is a very slow process. Humans tend to get tired by performing the same task again and again. Here, evaluation of sentences require great amount of consideration over several parameters. Based on these parameters a human judges the MT output. Moreover, an evaluator who is evaluating the output has to be paid for each sentence he has evaluated and in order to check the performance of an MT system; we generally test the system for several hundreds of sentences. So, this makes this process very costly (as compared to automatic evaluation).

Human judges often evaluate automatic translations on several different quality measures (like fluency, adequacy, compositionality etc). As a result this

process sometimes takes days to complete, which might delay the development of MT system.

2.1.2 Subjective

Human assessments are very subjective. On one hand, we may use multiple evaluators to test the same set of outputs, which always come up with different results. For example on a four point scale of fluency, one evaluator may give 2 (Little Comprehensible) to a translation whereas another evaluator may give 3 (Partially Comprehensible) to the same translation. On the other hand evaluators depend on evaluation guidelines involving several criteria which may vary from time to time or project to project thus may not be used twice.

2.2 Automatic Evaluation

In comparison to human evaluation, automatic evaluations are fast, inexpensive, objective and reusable. As and when demanded, automatic metrics provide an objective (numerical) score, which is a crucial aspect for MT developers in system development life cycle. In a broader perspective all the automatic metrics can be categorized as

- **Edit Distance Metrics:** where number of changes (insertion, deletions, and substitutions) required are counted, making it exactly like the reference sentence.
- **Precision Oriented Metrics:** where lexical matches are divided by total number of lexicons available in the output sentence.
- **Recall Oriented Metrics:** where lexical matches are divided by total number of lexicons available in the reference sentence.
- **F-measure Oriented Metrics:** where a collection of both precision and recall is used.

A large number of evaluation metrics has been proposed in the last decade which is based on either one of the categories and provides evaluation based on automatic reference translations. Papineni (2001) proposed BLEU (BiLingual Evaluation Understudy). This automatic evaluation metric has become the de-facto standard for all the MT Engine developers who need to quickly check the performance of their system as this metric employed the direct exploitation of reference translations. The score is evaluated by calculating the number of n-grams (word sequences) in the system output that are also present in the reference translations. A geometric mean of all such common chunks is calculated to generate the final score for the sentence. This is a precision oriented metric. This metric has recently been modified by Chen & Kuhn (2011) where they changed the calculating criteria from precision to recall. Snover at el (2006) proposed TER (Translation Edit Rate). This metric performed phrase reordering by allowing block movement of words (termed as shifts) within the MT output so as to make it exactly like reference translation. Banerjee & Lavie (2005) proposed METEOR (Metric for Evaluation of Translation with Explicit ORdering) which was an F-measure oriented metric. This metric was designed to

address the weaknesses of BLEU. In this metric, besides matching the lexicons of MT output and reference translations, stem and synonym matching was also done. The final score of the metric is calculated by harmonic mean. Lavie & Agarwal (2007) showed the higher correlation of the metric score with human judgments. In a more recent study, Denkowski and Lavie (2011) have extended this metric to paraphrase level. In this revision, they not only evaluated the lexicons but also the phrases of the MT output with reference translations. Leusch et al (2006) proposed CDER (Cover Disjoint Error Rate). This metric exploits the fact that the number of blocks in a sentence is equal to the number of gaps among the blocks plus one.

All these metrics are based on lexical similarities. These metrics have demonstrated notable quality to emulate human evaluators in different evaluation tasks. Working of these metrics is same and can be summarized by Figure 4. All these metrics, although same in working, differ in computation of lexical similarity.

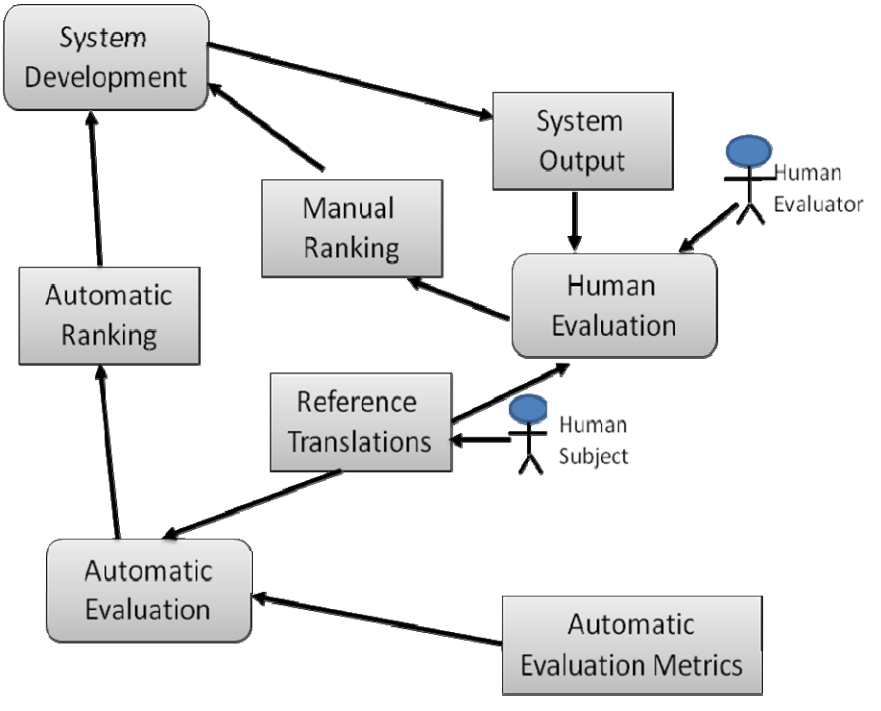


Fig. 4. Automatic Evaluation Process

We can correlate the results of automatic metric with that of human's subjective evaluation results. In this process, the output of the system can be provided to the human evaluator, who evaluates the output on the basis of a subjective

questionnaire/metric, based on which an evaluator can judge the output. Then, we can use either one or more automatic metrics. The metrics take one or more reference translations and match them with the output. Based on this criterion a score is computed. This process is then repeated for each sentence. Since automatic metric's score of an MT Engine is more objective, we can easily compare the results between different MT Engines or different versions of same MT Engine (to verify the increase in performance of an Engine in subsequent versions). We can also compare the results with human evaluation and verify the correctness of the automatic metrics.

3 Evaluation Methodology

We have created a test corpus of 600 sentences from Tourism Domain, which were uniformly distributed in a group of six documents of 100 sentences each. We registered the output of Google Machine Translation System, Anusaraka Language Assessor and an Example Based Machine Translation System (EBMT) developed at Banasthali (2011). For evaluation we used English-Hindi language pair. We performed human and automatic metrics for evaluation. For automatic evaluation we used BLEU and METEOR metrics. Each metric was evaluated at the sentence level, document level and system level.

Since BLEU was the first metric and is considered as de-facto metric, we employed it, as we wanted to study the implications of the same on to Hindi (a relatively free word order language). METEOR was used because this metric provided shallow linguistic features in evaluation, which helped in making somewhat precise judgements. Here we not only matched exact words but also matched morphological variants and synonyms. We developed and used a simple light weight stemming algorithm which was based on Rangnathan and Rao's stemmer (2003) and a synonym database.

For human evaluation, we created a 3 scale evaluation metric which addressed fluency and adequacy measures. The questionnaire so developed had eleven questions. Human evaluators were asked to answer all these eleven questions for each output sentence provided by each MT engine. The average score of each sentence, based on eleven inputs, was calculated. This process was repeated for all six documents.

4 Results

At first we performed analysis for document and system levels. We calculated the average score of each of the document's sentences as document's average score. Table 1 shows the results of this study. For all the documents, human evaluators and METEOR metric gave higher average score to Google. However, with BLEU metric, EBMT scored the highest average score for each document with only one exception in document two, where Anusaraka scored higher results.

Table 1. Document Level Scores of Engines

		Google	*EBMT*	*Anusaraka*
Document One	Human	**0.61**	0.09	0.22
	BLEU	0.34	**0.45**	0.37
	METEOR	**0.28**	0.05	0.19
Document Two	Human	**0.36**	0.16	0.06
	BLEU	0.40	0.40	**0.42**
	METEOR	**0.26**	0.05	0.19
Document Three	Human	**0.69**	0.05	0.20
	BLEU	0.33	**0.35**	0.33
	METEOR	**0.31**	0.07	0.20
Document Four	Human	**0.49**	0.09	0.21
	BLEU	0.33	**0.39**	0.26
	METEOR	**0.32**	0.06	0.18
Document Five	Human	**0.49**	0.14	0.28
	BLEU	0.32	**0.44**	0.34
	METEOR	**0.39**	0.07	0.26
Document Six	Human	**0.69**	0.13	0.3
	BLEU	0.25	**0.43**	0.26
	METEOR	**0.42**	0.07	0.26

At system level we took average score of each engine, metric wise. Table 2 shows the result of the same. Here again human and METEOR scores where highest for Google and BLEU score was highest for EBMT. Here, it is also clearly visible that human evaluation does not match with any of the automatic metrics. Since Google is the partially rule based MT engine its human and automatic scores had huge differences. This proves that lexical similarity based metrics tend to give high scores to statistical machine translation systems and thus are deemed to be engine biased. Figure 5 summarizes this data.

We also applied correlation on metrics at sentence and document levels. Since human evaluation metric is considered to be the golden metric, we applied correlation between human and automatic evaluation at sentence and document levels with all four engines. The results of the study are shown in Table 3.

Table 2. System Level Scores of Engines

	Human	*BLEU*	*METEOR*
Google	**0.5576**	0.3299	**0.3338**
EBMT	0.1091	**0.4130**	0.0609
Anusaraka	0.2120	0.3319	0.2160

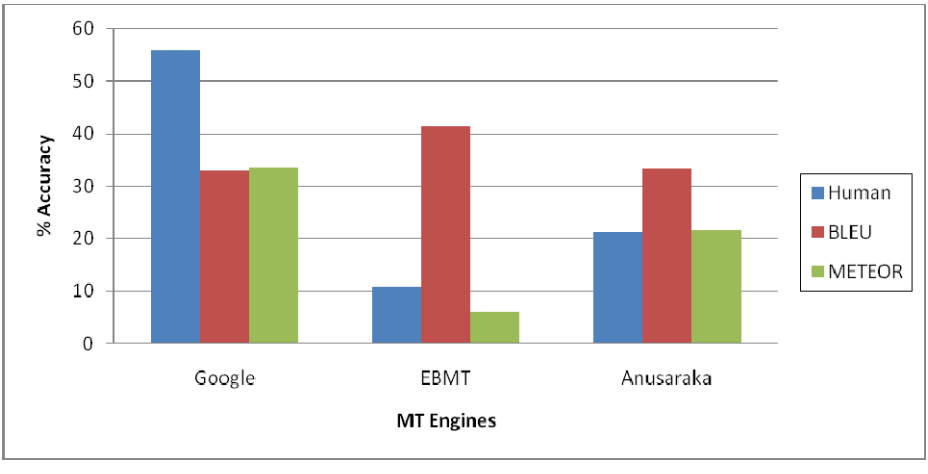


Fig. 5. Accuracy Achieved by Different Engines

Table 3. Correlation between Human and Automatic Metrics

		Human – BLEU	*Human – METEOR*
Google	Sentence Level	**0.052534**	0.036131
	Document Level	**0.010227**	**0.008683**
EBMT	Sentence Level	0.035987	0.00978
	Document Level	0.001975	0.000433
Anusaraka	Sentence Level	0.044701	**0.042383**
	Document Level	0.008623	0.00237

For sentence level, Google scored maximum in the Human-BLEU correlation while in Human-METEOR correlation, Anusaraka scored the highest. Table 3 summarizes the correlation between human and automatic metrics for all three MT Engines.

For document level, Google again showed the highest positive correlation between human metric and BLEU. At this level, Google also scored the highest correlation between human metric and METEOR.

5 Conclusion

We have shown evaluation results of three machine translators developed for English to Hindi Automatic Translation. We compared human evaluations with BLEU and METEOR automatic evaluation metrics. We found out that BLEU in several cases could not provide clear interpretations. It could be because their scoring mechanism is based on the assumption that all good translations of the same text would have similar translations. Unfortunately, due to the expressiveness and inherent ambiguity of the natural languages this assumption does not always hold good.

METEOR on the other hand produced good results, but still on several occasions failed to provide strong correlation with human evaluations. It may be because it works at shallow linguistic level. So, an immediate future study of this work could be to evaluate engine outputs at deeper linguistic levels. Another future study could be taken up to devise a mechanism to rank engine performances, either to rank different engines or to rank different version of same engine. This would help in better development of MT Engines.

References

Arnold, D., Balkan, L., Meijer, S., Humphreys, R., Sadler, L. (eds.): Machine Translation: An Introductory Guide. Blackwell-NCC, London (1994)

Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT Evaluation with improved correlation with human judgments. In: Proceedings Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and Summarization of Annual Meetings of Association of Computational Linguistics, pp. 65–72 (2005)

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: (Meta) Evaluation of machine translation. In: Proceedings of the Second Workshop on Statistical Machine Translation, pp. 136–158 (2007)

Chen, B., Kuhn, R.: AMBER: A modified BLEU, enhancing ranking metric. In: Proceedings of the Workshop on Statistical Machine Translation (2011)

Dabbadie, M., Hartley, A., King, M., Miller, K., Hadi, W.M.E., Popescu-Belis, A., Reeder, F., Vanni, M.: A Hands-On Study of Reliability and Coherence of Evaluation Metrics. In: Handbook of LREC 2002 Workshop Machine Translation Evaluation: Human Evaluation Meet Automated Metrics (2002)

Denkowski, D., Lavie, A.: Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In: Proceedings of the Workshop on Statistical Machine Translation (2011)

Falkedal, K.: Evaluation Methods for Machine Translation Systems: An Historical Survey and a Critical Account. ISSCO: Interim Report to Suissetra (1994)

Joshi, N., Mathur, I., Mathur, S.: Translation Memory for Indian Languages: An Aid for Human Translators. In: Proceedings of 2nd International Conference and Workshop in Emerging Trends in Technology (2011)

Lavie, A., Agarwal, A.: METEOR: an automatic metric for evaluation with high levels of correlation with human judgments. In: Workshop on Statistical Machine Translation at the 45th Annual Meeting of the Association of Computational Linguistics (2007)

Lehrberger, J., Bourbeau, L.: Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation. John Benjamin Publishers (1988)

Leusch, G., Ueffing, N., Ney, H.: CDER: Efficient MT Evaluation Using Block Movements. In: Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (2006)

Miller, G.A., Beebe-Center, J.G.: Some Psychological Methods for Evaluating the Quality of Translation. Mechanical Translations 3 (1956)

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation, RC22176. Technical Report, IBM T.J. Watson Research Center (2001)

Pfafflin, S.M.: Evaluation of Machine Translations by Reading Comprehension Tests and Subjective Judgments. Mechanical Translation and Computational Linguistics 8, 2–8 (1956)

Ramnathan, A., Rao, D.: A Lightweight Stemmer for Hindi. In: Proceedings of Workshop on Computational Linguistics for South Asian Languages, 10th Conference of the European Chapter of Association of Computational Linguistics, pp. 42–48 (2003)

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A Study of Translation Edit Rate with Targeted Human Annotation. In: Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA), pp. 223–231 (2006)

Wilks, Y.: Machine Translation: Its Scope and Limits. Springer, New York (2008)

A Novel Genetic Algorithm Based Method for Efficient QCA Circuit Design

Mohsen Kamrani[1], Hossein Khademolhosseini[2], Arman Roohi[2], and Poornik Aloustanimirmahalleh[3]

[1] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
kamrani@ce.sharif.edu
[2] Department of Computer Engineering, Science and Research Branch, Islamic Azad University, Tehran, Iran
{h.khademolhosseini,a.roohi}@srbiau.ac.ir
[3] Department of Electrical and Computer Engineering, Curtin University of Technology, Perth, WA 6845, Australia
poornik.aloustani@curtin.edu.au

Abstract. In this paper we have proposed an efficient method based on Genetic Algorithms (GAs) to design quantum cellular automata (QCA) circuits with minimum possible number of gates. The basic gates used to design these circuits are 2-input and 3-input NAND gates in addition to inverter gate. Due to use of these two types of NAND gates and their contradictory effects, a new fitness function has been defined. In addition, in this method we have used a type of mutation operator that can significantly help the GA to avoid local optima. The results show that the proposed approach is very efficient in deriving NAND based QCA designs.

Keywords: Genetic Algorithms, QCA, NAND gate, Hardware Reduction.

1 Introduction

In the recent decades vast researches on nanoscale have been studied to take the place of conventional transistor technology that have resulted in emerging technologies such as quantum cellular automata (QCA), single electron tunneling (SET) and tunneling phase logic (TPL).

Among these nanoscale devices, QCA could be of more interest due to its characteristics such as small dimension, low power consumption and high speed. A very important issue in the field of QCA circuitry is the optimization of logic gates employed in circuit design. To this end, many researchers have been done that try to design circuits with minimum possible required number of gates [1, 2].

In this paper, in order to design QCA circuits with efficient number of gates, a novel optimizer based on Genetic Algorithms (GAs) has been suggested. This optimizer designs the circuits by utilization of NAND gates and inverter gates.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 433–442.
springerlink.com

To find more efficient designs, two types of NAND gate (2-input and 3-input) have been used. The optimizer has been implemented in Python language and several experimental results have been verified by QCADesigner.

2 Background

2.1 GA Review

Genetic Algorithm (GA) [3] as a branch of Evolutionary Algorithms is an optimization tool that simulates the natural selection process to find solutions to the problems. The efficiency of the GAs in solving combinatorial optimization problems in addition to their intrinsic ability in optimization of objective functions irrespective of the gradient or higher derivatives of them has made them a popular tool for solving many combinatorial problems including the logic optimization problem of QCA circuits. The main body of a typical GA that has been used in this paper is explained in the following paragraphs.

GA begins the search procedure by creating an initial set of some randomly selected solutions (chromosomes) to the problem, called initial population. Afterwards GA evaluates each chromosome to lead the population in a proper way to the optimum solutions. Evaluation of the chromosomes is carried out based on a predefined function called fitness function that assigns a value to each chromosome according to how far or close it is from the optimum point.

After evaluation of the chromosomes, some of them should be selected based on a probability proportional to their fitness in order to generate the next generation. The selected chromosomes are divided into three parts that generate the next population. The first part includes some of the chromosomes having the highest fitness values also called Elite chromosomes that move directly to the next generation. The remaining parts of the selected chromosomes are utilized by the crossover operator and the mutation operator.

The crossover operator generates one or more new offspring(s) from the selected parents to achieve new solutions. There are many types of crossover that can be used based on the presentation of the chromosomes and some other aspects. For instance, sub-tree crossover is a method that can be employed when the chromosomes are represented by trees. This method applies on two chromosomes and exchanges a sub tree of the first chromosome with a sub tree of the second one to generate a child.

The mutation operator is another tool which can help GA to escape the local optima and achieve the global optima. Mutation operator usually applies on a chromosome to slightly change it so as to generate a child.

Finally the routine should be repeated until a termination condition satisfies, which means that GA reaches the maximum number of generations. More information about GA is provided in [4-8].

2.2 Review of QCA

Quantum-dot cellular automata (QCA) is a new nanoscale technology which functions based on Coulombic interaction instead of current used in conventional CMOS. QCA has attractive features like high speed operation and small dimension. In QCA

technology binary information is encoded by formations of electrons. Thus, power consumption, a major obstacle in VLSI designs, in QCA circuitry is low which is due to current-less feature. The basic structure in QCA is a cell which consists of four dots and two identical electrons. Each dot can be occupied by one of the two hopping electrons. On account of the dynamic behavior of the electrons, they arrange themselves diagonally in order to reach to the maximum distance. As shown in Fig. 1 two possible polarizations might occur, which represent the binary values "0" and "1" [9-11].

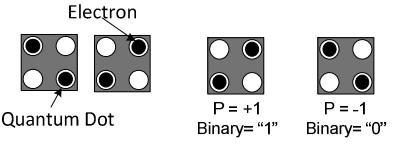


Fig. 1. QCA cell and two possible polarizations

A basic structure which can be constructed by an array of aligned cells is a QCA wire, Fig. 2(a). Two other primary building blocks which are the base of many QCA designs are the inverter and the majority gate [12-16]. The inverter is made up of four QCA wires, as shown in Fig. 2(b) [10].

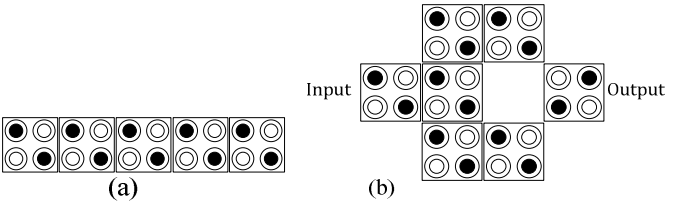


Fig. 2. (a) A QCA wire, (b) A QCA inverter

A Three-input [10] and a five-input majority gates [12, 18] are illustrated in Fig. 3(a) and Fig. 3(b), respectively. The 3-input majority gate acts according to the following equation:

$$M(A,B,C) = AB + AB + BC\,. \tag{1}$$

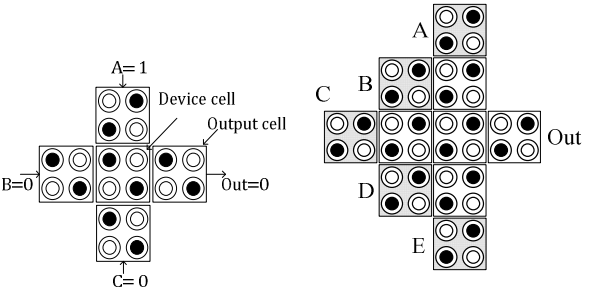


Fig. 3. (a) A 3-input majority gate, (b) A 5-input majority gate

The 5-input majority gate functions as (2). Three (Five) input Majority gate can be programmed as AND or OR gate by fixing one (two) of input cells to p = -1 or p = +1, respectively [10, 12].

$$M(A,B,C,D,E) = \text{ABC+ABD+ABE+ACD+ACE} \\ \text{+ADE+BCD+BCE+BDE+CDE}\,. \tag{2}$$

Due to the advantages of the NAND gate such as functional completeness, this work will concentrate on NAND-based designs to implement QCA circuits. The majority gate can be configured to act as a NAND gate which is chosen as the fundamental logic block. In fact, it is inferable that most of the Boolean logic circuits in QCA can be realized using only QCA NAND gate (Fig. 4). Several synthesis algorithms are available to convert a Boolean logic function to a NAND gate logic function.

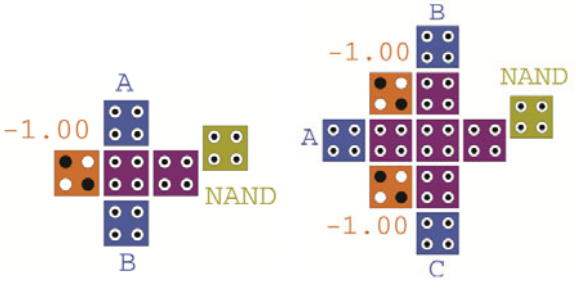


Fig. 4. Layout (Schematic) of 2-input NAND and 3-input NAND

3 Proposed Approach

In this paper, we have proposed a method to automatically design QCA logic circuits which are implemented by combination of 2-input and 3-input NAND gates in addition to the inverter gate. The GA has been utilized in this method to optimize the number of gates and clock cycles. In the following, we have explained the specifications of the proposed method.

As a natural choice a tree structure has been utilized in order to represent the chromosomes. In this tree structure the root and the inner nodes of the tree are either a NAND gate specified with the 'Nd.' or an inverter gate specified with the 'Inv.' and the leaves of the tree are the inputs of the circuit or fixed cells. The inputs to each gate (placed in the inner nodes) are specified with the branches rooted at that node. This structure is illustrated with an example in Fig. 5 that shows the representation of N(N(A',B),1,N(A,C')).

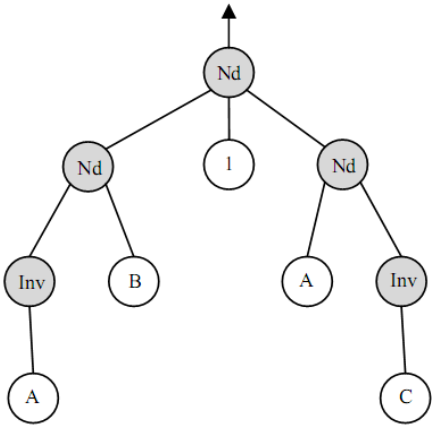


Fig. 5. The chromosome that represents N(N(A',B),1,N(A,C'))

The algorithm begins with initialization of the population with n randomly generated chromosomes, where n is an integer between 300 and 1000, specified manually depending on the complexity of the circuit.

To evaluate the chromosomes the following fitness function has been defined:

$$f(c_i) = \frac{N(p, c_i)}{|p|} + \frac{1}{N(q, c_i) + Nodes(c_i)} \tag{3}$$

In the above formula, N(.,.) denotes the function that calculates the number of minterms given as the first parameter implemented by the second parameter, p contains the minterms to be implemented, |p| is the size of m that has been used for scaling issues and q denotes the rest of minterms that should not be implemented by the circuit. To declare a fitness function for this problem the positive and negative effects of utilizing the 3-input NAND gate should be taken into account. The positive effect is that these gates can lead to reduction of the levels of the circuits. The negative effect of these gates is that these gates can be implemented with more cells compared to 2-input NAND gates that might increase the total cell number. Hence in order to consider both of these effects, the 3-input NAND gates have been counted as 1.6 nodes in the Nodes function while the 2-input NAND gates and inverter gates have been counted as 1 node in the Nodes function.

The tournament selection has been used in order to select the parents for reproduction operators. This operator selects each chromosome based on its rank achieved by the fitness value assigned to that chromosome in the evaluation phase of the algorithm.

As mentioned in the previous section, the sub-tree crossover has been utilized as the crossover operator. The operator is exemplified in Fig. 6 that illustrates the combination of two chromosomes which represent N(A,B,C) and N(N(B',A),C') from left to right, respectively.

Mutation of a chromosome has been carried out by recombination of a randomly generated chromosome and a parent chromosome to achieve a child chromosome.

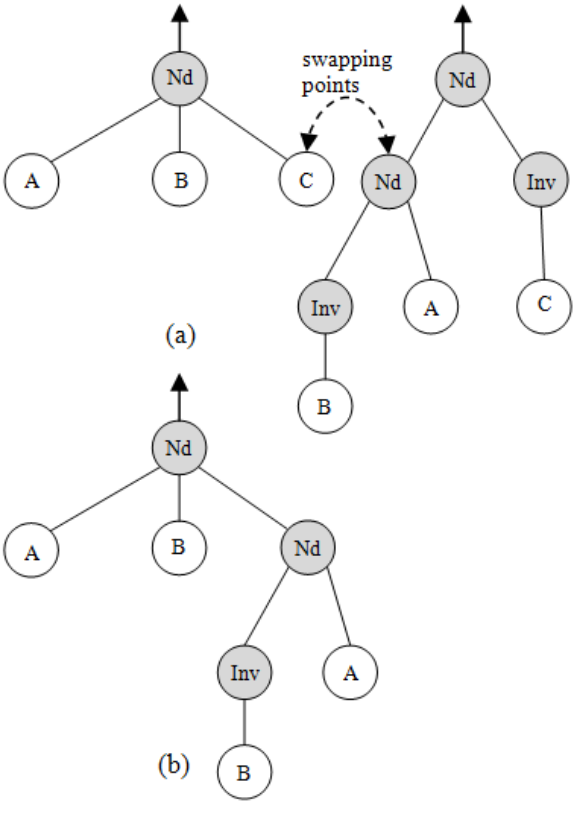


Fig. 6. (a) Two parent chromosomes before crossover, and (b) a child as the result of the crossover operation.

The algorithm stops when total number of generations exceeds maxGen (a number between 1000 and 2000), or no improvement in fitness function happens during more than fLim (a number between 20 and 30) generations.

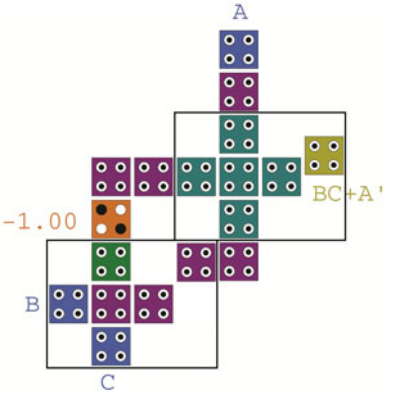
4 Experimental Results

To study the effectiveness of the proposed NAND gate-based circuit optimization method, the method has been implemented in Python and some of the results are verified by QCADesigner tool. The experiments were done on a 2.2 GHz Intel Pentium IV machine with 512MB RAM running Windows XP Professional and the results have been shown in TABLE 1.

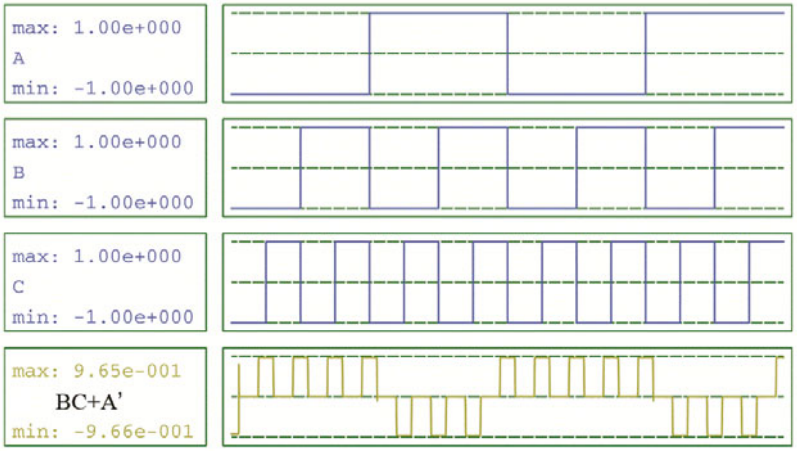
The functionality verification of the proposed gate and circuit are carried out using the QCADesigner bistable engine [17]. The following parameters are used for a bistable approximation: cell size = 18nm, number of samples = 50000, convergence tolerance = 0.0000100, radius of effect = 65.000000nm, relative permittivity = 12.900000, clock high = 9.800000e−022 J, clock low = 3.800000e−023 J, clock shift = 0, Clock

Table 1. Optimization of Some Standard Functions.

No.	Minterms	NAND-based implementation
1	m_7	N(A,B,C)′
2	$m_0+m_1+m_2+m_3+m_7$	N(A,N(B,C))
3	$m_1+m_3+m_5+m_6+m_7$	N(C′,N(A,B))
4	m_1+m_4	N(N(C′,B′,A),N(A′,B′,C))
5	$m_5+m_6+m_7$	N(N(C,B),N(C,A))
6	$m_0+m_2+m_3+m_4+m_5+m_6+m_7$	N(B,A,C′)
7	$m_1+m_3+m_4+m_6$	N(N(C′,A),N(A′,C))



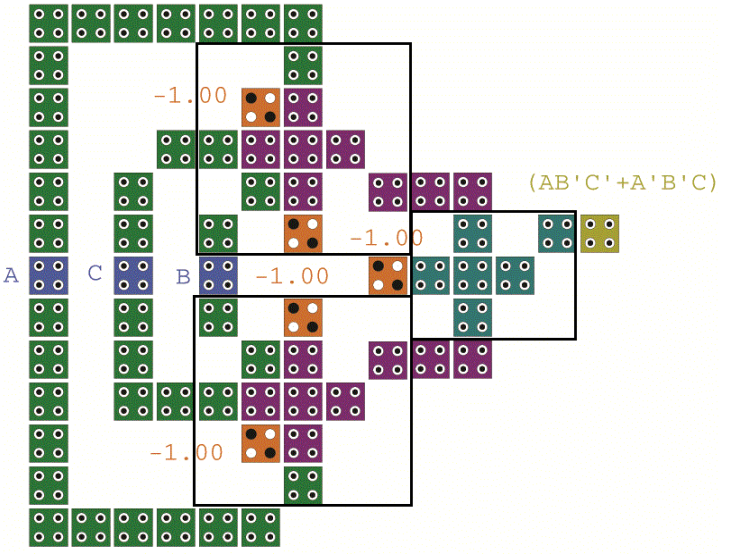
(a)



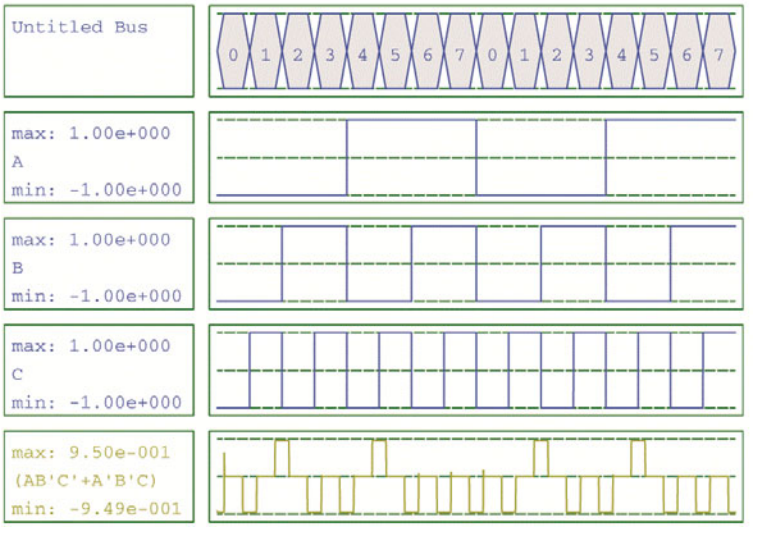
(b)

Fig. 7. (a) Schematic for minterms $\sum m(0,1,2,3,7)$, (b) Simulation result for $\sum m(0,1,2,3,7)$

amplitude factor = 2.000000, layer separation = 11.500000, maximum iterations per sample = 100. Most of the mentioned parameters are default values in QCADesigner. Two functions are generated using the proposed method and their schematics accompanying by the simulation results are depicted in Fig. 7 and Fig. 8.



(a)



(b)

Fig. 8. (a) QCA schematic for $\Sigma m(1,4)$, (b) Simulation result for $\Sigma m(1,4)$

5 Conclusion

The design of circuits having efficient numbers of cells and levels in their structures implemented with QCA is very important. In this paper we have proposed a method based on GA to reduce the number of cells required to design a QCA circuit composed of 2-input and 3-input NAND gates in addition to the inverter gates. A new fitness function has been proposed that takes the positive and negative effects of the combination of 2-input and 3-input NAND gates into account. Finally some samples were simulated using QCADesigner tool. The results show that this method produces reasonable designs with minimum possible number of gates which are extensible to designs based on other fundamental QCA elements.

References

1. Bonyadi, M.R., et al.: Logic Optimization for Majority Gate-Based Nanoelectronic Circuits Based on Genetic Algorithm. In: International Conference on Electrical Engineering, ICEE 2007, pp. 1–5 (2007)
2. Zhi, H., Qishan, Z., Haruehanroengra, S., Wei, W.: Logic optimization for majority gate-based nanoelectronic circuits. In: Proceedings of 2006 IEEE International Symposium on Circuits and Systems, ISCAS 2006, p. 4, p. 1310 (2006)
3. Holland, J.H.: Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. University of Michigan Press, Ann Arbor (1975)
4. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Publishing Company Inc., Massachusetts (1989)
5. Davis, L.: Handbook of Genetic Algorithms. Van Nostrand Reinhold, New York (1991)
6. Chambers, L.D.: Practical Handbook of Genetic Algorithms: New Frontiers, 1st edn. CRC Press, Inc., Boca Raton (1995)
7. Chambers, L.D.: Practical Handbook of Genetic Algorithms: Complex coding system, 1st edn. CRC Press, Inc., Boca Raton (1998)
8. Chambers, L.D.: Practical Handbook of Genetic Algorithms: Applications, 2nd edn. CRC Press, Inc., Boca Raton (2000)
9. Lent, C.S., Tougaw, P.D., Porod, W., Bernstein, G.H.: Quantum cellular automata. Nanotechnology 4(1), 49–57 (1993)
10. Tougaw, P.D., Lent, C.S.: Logical devices implemented using quantum cellular automata. Journal of Applied Physics 75(3), 1818–1825 (1994)
11. Lent, C.S., Tougaw, P.D.: A device architecture for computing with quantum dots. Proceedings of the IEEE 85(4), 541–557 (1997)
12. Navi, K., Sayedsalehi, S., Farazkish, R., Rahimi Azghadi, M.: Five-input majority gate, a new device for quantum-dot cellular automata. Journal of Computational and Theoretical Nanoscience 7, 1546–1553 (2010)
13. Sayedsalehi, S., Moaiyeri, M.H., Navi, K.: Novel Efficient Adder Circuits for Quantum-Dot Cellular Automata. To be Published in Journal of Computational and Theoretical Nanoscience (2011)
14. Cho, H., Swartzlander, E.E.: Adder and multiplier design in quantum-dot cellular automata. IEEE Transactions on Computers 58(6), 721–727 (2009)

15. Navi, K., Farazkish, R., Sayedsalehi, S., Rahimi Azghadi, M.: A new quantum-dot cellular automata full-adder. Microelectronics Journal 41, 820–826 (2010)
16. Roohi, A., Kamrani, M., Sayedsalehi, S., Navi, K.: A Combinational Logic Optimization for Majority Gate-Based Nanoelectronic Circuits Based on GA. In: International Semiconductor Device Research Symposium, USA (2011)
17. QCADesigner Home Page, http://www.atips.ca/projects/qcadesigner
18. Roohi, A., Kamrani, M., Khademolhosseini, H., Sayedsalehi, S.: A Novel Symmetric Design for 5-input Majority Gate in Quantum Cellular Automata. In: International Conference on NanoScience, Engineering and Technology, ICONSET 2011 (2011)

Adaptation of Cognitive Psychological Framework as Knowledge Explication Strategy

S. Maria Wenisch[1], A. Ramachandran[2], and G.V. Uma[3]

[1] Department of Information Science and Technology, Anna University, Chennai 600 025, India
wenischs@gmail.com
[2] Center for Climate Change and Adaptation Research, Anna University,
Chennai 600 025, India
ram7@annauniv.edu
[3] Department of Information Science and Technology, Anna University, Chennai 600 025, India
gvuma@annauniv.edu

Abstract. The differences between the way an indigenous expert understands and the way a scientific expert understands a natural resource which are incomplete in themselves can become complementary and become a major strength in achieving sustainability. An integrated system approach is the best practice of managing natural resource. In knowledge management integration is viewed in terms of horizontal and vertical dimensions. The study presents the possibility of knowledge sharing and integration between indigenous and scientific experts in a multi level multi criteria decision making environment using a cognitive psychological model of knowledge discovery called Johari Window model for knowledge sharing in the management of natural resource. Knowledge integration facilitates higher level of knowledge explication. Johari window presents a framework for knowledge integration. The advantage of this model is that it takes the problem of 'the fourth quadrant', where very large totally unexplored unpredicted outliers lie, into its account. Knowledge system becomes more robust, efficient and sustainable by narrowing down the knowledge gap between the experts.

1 Introduction

Knowledge elicitation and sharing among the stakeholders in natural resource management system is one of the key issues. Knowledge on natural resource is either experiential gained by using the natural resource in a particular location or experimental gained by scientific experiments performed on the natural resource. Experiential is a bottom up approach because the subject uses a natural resource and gains knowledge of it in the process of using it. Experimental takes a top down approach. Based on a proposition, one tries to fit the natural resource into that framework of the proposition. Indigenous knowledge is mostly experiential, knowledge gained out of constant familiarity with the reality. Rekha Singhal states that

> Indigenous refers to knowledge and practices that have originated locally and are performed by a community or society in a specific place. This knowledge evolves and emerges continually over time according to peoples perception

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 443–450.
springerlink.com

> and experience of their environment and is usually transmitted from generation to generation by word of mouth or by practice. In contrast, scientific forestry utilizes specialized knowledge for managing forest resources not only for local populations but also for wider objectives and the global scientific forestry community. Scientific knowledge on forest management is generally shared in formal, written, and non-traditional ways.

Scientific knowledge is mostly experimental, knowledge gained out of rigorous study based on a proposed theory and experimenting with prototypes. The best method for natural resource management is to combine the experiential and experimental knowledge. Ataur Rahman has enumerated some distinctions between traditional and scientific knowledge systems and has attached explicit nature to scientific knowledge and tacit nature to indigenous knowledge. Agarwal has insisted that

> there is a need to move beyond the dichotomy of indigenous versus scientific and work towards building bridges across the indigenous and scientific divide.

Focusing on the development of the framework for a methodological integration of indigenous and scientific knowledge Jessica Mercer et al have worked on the integration framework. Johan et al have suggested that knowledge integration processes may benefit from early recognition of the dualities at hand and strategies aimed at creating thirdness, including some suggestions on the concrete forms such thirdness may take. Stefano et al have argued that an autonomy and experimental climate (i.e. shared perception that the team supports autonomous action and experimentation and risk taking) can favor the teams ability to integrate members knowledge. Chen Kun et al have proposed a general knowledge mediation infrastructure for multi-agent systems. Hsiu-Ling et al have suggested that firms should be cautious in their pursuit of a strategy of vertical integration, given the non-monotonic impact on innovative performance, whilst an increase in the level of vertical integration is also likely to diminish the effectiveness of the external knowledge sourcing. Ad Breukel et al's findings enable entrepreneurs to participate in efforts to enhance their ICT capabilities and moderation of effective knowledge sharing within a destination and event marketing platform. Rekha Singhal would argue that

> there is no fixed method of addressing the bottlenecks in integration of indigenous and scientific knowledge, instead the methods chosen will vary according to what is appropriate and feasible within the institutional, ecological, and social environments in which they operate.

For example, Rist et al have worked on the role of ethosciences in the dialogue between western scientific knowledge and indigenous scientific knowledge. This paper proposes Johari Window as the model for knowledge sharing between indigenous and scientific natural resource management experts. Johari algorithm is proposed for combining the indigenous and scientific knowledge of natural resource.

2 Proposed Methodology

Johari Window is a framework developed and by Joseph Luft and Harry Ingham as a cognitive psychological approach to self discovery. The model can also be utilized for

team building and group interactions for attaining a higher level of understanding between the members. The Johari Window model consists of a fours quadrants (1).Johari Window is used for representing the knowledge on a particular entity held by indigenous and scientific experts. You and Me of Johari Window categories are used to represent indigenous and scientific experts respectively. The four quadrants are: Quadrant 1 represents the knowledge held by both the indigenous expert (E_i) and scientific expert (E_s) on the entity concerned. Quadrant 2 represents the state where knowledge is held by the E_i alone. Quadrant 3 represents the state in which E_s alone holds the knowledge on the entity under consideration. The advantage of Johari Window model is that it incorporates the black spot (unknown) into its system of representation.

Table 1. Four Quadrants

Quadrants	Property
Quadrant1	Known by You and Me
Quadrant2	Known by Me
Quadrant3	Known by You
Quadrant4	Unknown to both

Table 2. Indigenous and Scientific Experts in Johari Window

Known by E_i and E_s	Known by E_i alone
Known by E_s alone	Unknown to both

Table 3. Binary Representation of Four States

Q_{11}	Q_{10}
Q_{01}	Q_{00}

The knowledge states of the four quadrants are represented using binary suffixes. The four Johari quadrants are identified by Johari variable k_i (indigenous knowledge) and k_s (scientific knowledge). The four quadrants of Johari window has four combinations of knowledge states: 00, 10, 01, and 11. The knowledge states are either independent or dependent states. A dependency presupposes proceeding and succeeding knowledge states. For example if knowledge state $K3$ is dependent on $K2$ and $K2$ is dependent on $K1$, it automatically holds transitive property. Thus we can call Q_{11}, Q_{10}, Q_{01}, and Q_{00} as the four knowledge states between two knowledge holders. Knowledge sharing is possible when the knowledge of a particular property of an object is either in quadrant 2 or quadrant 3. The mutual sharing of knowledge between quadrant 2 and quadrant 3 would result to quadrant 1. Two possibilities of a particular knowledge reaching quadrant 1 are

$$Q_{10} \rightarrow Q_{01} \rightarrow Q_{11} \tag{1}$$

$$Q_{01} \rightarrow Q_{10} \rightarrow Q_{11} \tag{2}$$

3 The Problem of Dark Spot

Let us consider w_i and w_s as the worlds of indigenous and scientific experts respectively. Let p_i and $p_s$be the properties of an object O in study known to the indigenous world w_i and the scientific world w_s respectively. And let p_{xy} represent the properties with two index where xy are the combination of indigenous and scientific experts. The 2^n possible states are: 01,10,11,00. In Johari window analysis the four states represent four quadrants. Then there may be Q_{00},the fourth quadrant which is the dark spot for the actors in the knowledge sharing network as it lies outside the possible worlds w_i and w_s of indigenous and scientific experts respectively. Johari framework proposes that the fourth quadrant is to be minimized by frequent sharing of the knowledge between the experts and making a cumulative world w_{is} by combining w_i and w_s. The integration of two worlds is given as

$$w_{is} = w_i \cap w_s \tag{3}$$

Thus when there are n experts involved in knowledge sharing the integrated knowledge of world

$$w_{is} = [w_{i1} + w_{i2} \ldots w_{in}] \bigcap [w_{s1} + w_{s2} \ldots\ldots\ldots w_{sn}] \tag{4}$$

$$w_{ds} = (w_i \cup w_s) - (w_i \cap w_s) \tag{5}$$

Let us consider w_{ds} the dark spot shown in figure 1. A dark spot can be any knowledge that is not a member of w_i or w_s. Let k be a knowledge component and w_i and w_s be worlds of indigenous and scientific experts respectively. If $k \in w_i$ then $k \rightarrow k_i$. If $k \in w_s$ then $k \rightarrow k_s$. If $k \in w_i \wedge w_s$ then $k \rightarrow k_{is}$. If $k \notin k_i \vee k_s$ then $k \in k_{ds}$ and in this case k would belong to the list to be updated by E_i and E_s.

The discovery of new k by E_i or by E_s might widen the world of w_{is}. In this case we have to decide on the nature of the new k. The new k may either add to k_{is} or may distort or contradict or falsify the existing k_{is}. If the impact factor of the new k

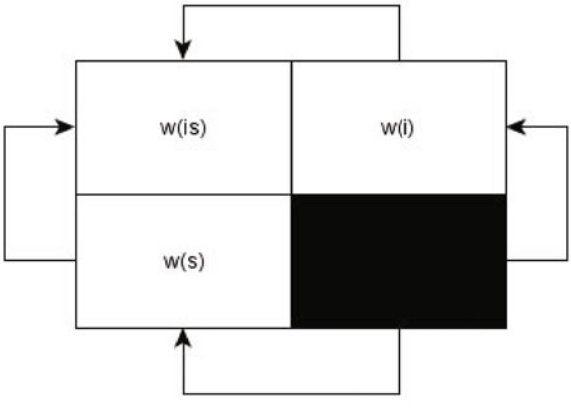


Fig. 1. Unexplored Dark Spot illustrated using Johari Window Framework

is significant enough to modify the existing k_{is}, then we can identify it as a $-k$ or a $+k$. Depending on the nature of k the new knowledge either $(+k)$ adds to the existing component of knowledge or $(-k)$ modifies or falsifies the existing component of knowledge.

4 Johari Algorithm

Let us consider the following table 4 and we call it a Johari table which contains object, properties and K_i flag and K_s flag. If the properties of the objects are known to indigenous expert then K_i flag is set else it is unset. Similarly if the properties of the objects are known to scientific expert then the flag is sent else it is unset. The algorithm stores properties in the property table with K_i and K_s flags set or unset. Table.2, for example K_i of p_1 of O_1 is set while K_s of the same property is unset. Similarly K_i of p_2 of O_1 is unset while K_s of the same property is set. It means that p_1 of O_1 is from the Knowledge base of indigenous expert and p_2 of O_1 is from the knowledge base of scientific expert. If p_1 of O_1 is shared with the scientific expert then K_s flag of p_2 of O_1 will be set. In the same way, if p_2 of O_1 is shared with indigenous expert, then K_i flag of p_2 of O_1 will be set. Now p_1 of O_1 and p_2 of O_1 would belong to quadrant1 Q_{11}.

The Johari algorithm reads the list of the properties of an object under consideration. Let p_i be a property of an object O_i. The algorithm checks the source of the property and then sets the flag according to the source. The flag K_i is set if p_i is from an indigenous expert E_i or the flag K_s is set if p_i is from a scientific expert E_s. Otherwise we know that it is neither known to K_i nor K_s and so it belongs to quadrant Q_{00}. If the learning set of indigenous expert and scientific expert be S_i and S_s respectively. A Johari search algorithm can search a database DB_{is} and classify the S_i, S_s and S_{ds} sets where S_i contains k_i from E_i and S_s contains k_i from E_s and S_{ds} contains k which does not belong either to S_i or S_s.

$$p_i \in S_i \wedge p_i \in S_s \rightarrow p_i(known) \tag{6}$$

$$p_i \in S_i \wedge p_i \notin S_s \rightarrow move(p_i, learningset(S_s)) \tag{7}$$

$$p_i \notin S_i \wedge pi \in S_s \rightarrow move(p_i, learningset(S_i)) \tag{8}$$

$$p_i \notin S_i \wedge p_i \notin S_s \rightarrow move(p_i, (learningset(S_i) \wedge learningset(S_s))) \tag{9}$$

From the above equations (6), (7), (8), (9) Johari Learning set algorithm is arrived for finding the learning sets for indigenous and scientific experts.

Table 4. Property Table with K_i and K_s flags

Object	Property	K_i	K_s
O_1	p_1	Yes	No
O_1	p_2	No	Yes
O_2	p_1	Yes	Yes

Algorithm: Johari Classification

$L-i$: List of properties of K_i
L_s : List of properties of K_s
p_i :Property of O_i
O_i : Object
DB_{is} : Integrated Database
Read p_i of O_i from L_i
Read p_i of O_i from L_s
If $p_i \in K_i \wedge p_i \in K_s$ then
Store p_i in DB_{is}
Set K_i and K_s
else
If $p_i \in K_i \wedge p_i \notin K_s$ then
Set K_s
else
If $p_i \notin K_i \wedge p_i \in K_s$ then
Set K_s
else
Unset $K_i \wedge K_s$

In the figure 2,the states S_0 and S_f represent the initial and final states of the knowledge flow respectively. The states S_1, S_2, S_3, S_4 represent the four intermediate states.The thick lines represent the flow to the final state. The final state S_f can be reached only through the intermediate states S_1, S_2, S_3, S_4. There is knowledge flow between the intermediate states. When the flow does not take place between any one of the intermediate states, the knowledge may remain incomplete. From the equations 1 and 2 the initial, intermediate, and final states can be represented in table 5

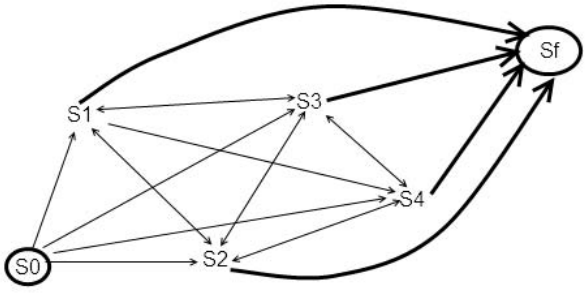


Fig. 2. Initial State, Intermediate States and Final State of Knowledge Flow

Table 5. States Involving Four Experts

Level	Initial States	Intermediate States	Final States
1	0000	-	-
2	0000	0001,0010,0100,1000	-
3	0001,0010,0100,1000	0011,0110,1100	-
4	0011,0110,1100	0111,1011,1101,1110	-
5	0111,1011,1101,1110	-	1111

Algorithm: Creating Johari Learning Sets

DB_{is} : Integrated Database
K_i; K_s : Flags
S_i; S_s : Learning Sets
Read DB_{is}
If K_i flag and K_s flag NOT set then
add $p_i to S_i \wedge S_s$
else If K_i flag is NOT set then
add p_i to S_i
else If K_s flag is NOT set then
add p_i to S_s

5 Conclusion

Theoretical analysis of the framework for the knowledge integration of the indigenous and scientific experts using Johari Window, the cognitive psychological model has proved to address the problem of the fourth quadrant effectively. The model facilitates the explication of unknown dark spot in knowledge integration scenario. The explication process has been identified to be operative in different states of integration. The number of the experts involved in integration process linearly increases the quantity and the quality of explication process.The uniqueness of this framework lies in its 'dialogical model'. The knowledge system becomes more robust and efficient by achieving a greater degree of knowledge explication.

Theoretical cognitive psychological framework can be further studied and applied in a particular domain. In a typical natural resource management scenario, the implementation of the framework needs a designing techniques for acquisition, creation, and integration of indigenous and scientific knowledge. The representation of knowledge of indigenous and scientific experts are another major area of exploration.

References

1. Agrawal, A. : Indigenous and Scientific Knowledge: Some Critical Comments. IK Monitor 3 (2004), http://www.nuffic.nl/ciran/ikdm/3-3/articles/agrawal.html (cited January 5, 2004)

2. Rahman, A.: Development of an Integrated Traditional and Scientific Knowledge Base: A Mechanism for Accessing, Benefit-Sharing and Documenting Traditional Knowledge for Sustainable Socio-Economic Development and Poverty Alleviation. UNCTAD, Geneva (2000)
3. Rotha, C., Cointet, J.-P.: Social and Semantic Coevolution in Knowledge Networks. Social Networks 32, 16–29 (2010)
4. Chen, K., Wang, H., Lai, H.: A General Knowledge Mediation Infrastructure for Multiagent Systems. Expert Systems with Applications 38, 495–503 (2011)
5. Yang, C.-W., Fang, S.-C., Lin, J.L.: Organisational Knowledge Creation Strategies: A Conceptual Framework. International Journal of Information Management 30, 231–238 (2010)
6. Lwoga, E.T., Ngulube, P., Stilwell, C.: Managing Indigenous Knowledge for Sustainable Agricultural Development in Developing Countries: Knowledge Management Approaches in the Social Context. The International Information & Library Review 42, 174–185 (2010)
7. Chang, H.H., Chuang, S.-S.: Social Capital and Individual Motivations on Knowledge Sharing:Participant Involvement as a Moderator. Information & Management 48, 9–18 (2011)
8. Li, H.-L., Tang, M.-J.: Vertical Integration and Innovative Performance:The Effects of External Knowledge Sourcing Modes. Technovation 30, 401–410 (2010)
9. Mercer, J., Kelman, I., Suchet-Pearson, S., Lloyd, K.: Integrating Indigenous and Scientific Knowledge Bases for Disaster Risk Reduction in Paupa New Guinea. Geografiska Annaler: Series B, Human Geography 91(2), 157–183 (2009)
10. Mu, J., Tang, F., MacLachlan, D.L.: Absorptive and Disseminative Capacity: Knowledge Transfer in Intra-organization Networks. Expert Systems with Applications 37, 31–38 (2010)
11. Hovelynck, J., Dewulf, A., Francois, G., Taillie, T.: Interdisciplinary Knowledge Integration Through Group Model Building: Recognizing Dualities and Triadizing the Conversation. Environmental Science & Policy 13, 582–591 (2010)
12. Janhonen, M., Johanson, J.-E.: Role of Knowledge Conversion and Social Networks in Team Performance. International Journal of Information Management 31(3), 217–225 (2010)
13. Singhal, R.: A Model for Integrating Indigenous and Scientific Forest Management: Potentials and Limitations for Adaptive Learning. In: Lawrence, A. (ed.) Forestry. Forest Users and Research: New Ways of Learning, pp. 131–137. ETFRN, Wageningen (2000)
14. Basaglia, S., Caporarello, L., Magnib, M., Pennarola, F.: IT Knowledge Integration Capability and Team Performance: The Role of Team Climate. International Journal of Information Management 30, 542–551 (2010)
15. Rist, S., Dahdouh-Guebas, F.: Ethnosciences–A Step Towards the Integration of Scientific and Indigenous Forms of Knowledge in the Management of Natural Resources for the Future. Environ. Dev. Sustain (2006), doi:10.1007/s10668-006-9050-7
16. Wang, X., Stolein, M., Wang, K.: Designing Knowledge Chain Networks in China — A proposal for a Risk Management System Using Linguistic Decision Making. Technological Forecasting & Social Change 77, 902–915 (2010)
17. Li, Y.-M., Jhang-Li, J.-H.: Knowledge Sharing in Communities of Practice: A Game Theoretic Analysis. European Journal of Operational Research 207, 1052–1064 (2010)

Comparing Fuzzy-C Means and K-Means Clustering Techniques: A Comprehensive Study

Sandeep Panda, Sanat Sahu, Pradeep Jena, and Subhagata Chattopadhyay

Dept. of Computer Science and Engineering
National Institute of Science and Technology
Palur Hills, Berhampur 761008 Odisha India
{sandeeppandakumar15081991,sanat.lipu,
subhagatachatterjee}@gmail.com, pradeep1_nist@yahoo.com

Abstract. Clustering techniques are unsupervised learning methods of grouping similar from dissimilar data types. Therefore, these are popular for various data mining and pattern recognition purposes. However, their performances are data dependent. Thus, choosing right clustering technique for a given dataset is a research challenge. In this paper, we have tested the performances of a Soft clustering (e.g., Fuzzy C means or FCM) and a Hard clustering technique (e.g., K-means or KM) on Iris (150 x 4); Wine (178 x 13) and Lens (24 x 4) datasets. Distance measure is the heart of any clustering algorithm to compute the similarity between any two data. Two distance measures such as Manhattan (MH) and Euclidean (ED) are used to note how these influence the overall clustering performance. The performance has been compared based on seven parameters: (i) sensitivity, (ii) specificity, (iii) precision, (iv) accuracy, (v) run time, (vi) average intra cluster distance (i.e. compactness of the clusters) and (vii) inter cluster distance (i.e. distinctiveness of the clusters). Based on the experimental results, the paper concludes that both KM and FCM have performed well. However, KM outperforms FCM in terms of speed. FCM-MH combination produces most compact clusters, while KM-ED yields most distinct clusters.

Keywords: Clustering, FCM, KM, Distance measures, Performance test.

1 Introduction

Clustering is a method of grouping similar data and distinctly separating them from the dissimilar data. It helps recognizing hidden patterns within the data. It is an unsupervised approach. For pattern extraction, clustering techniques depend on the similarity measures between the representative and the data to be clustered. Representative data denotes the cluster center, i.e., the ideal data of the cluster. Similarity is computed based on the distance measure between the cluster center and the data to be clustered using several methods, such as Manhattan, Euclidean, Cosine, Mahalanabis, and Hamming etc. The advantages of clustering techniques are that these do not require domain knowledge and labeled data, are able to deal with various types of data (including noisy data and outliers), capable of interpreting ad-hoc data and could be reused.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 451–460.
springerlink.com

There are two broad types of clustering methods, e.g., 'Soft' and 'Hard' clustering. Soft clusters are devoid of distinct boundaries, as seen in Fuzzy C Means (FCM) [1], Fuzzy K-nearest Neighbor (FKN) [2], Entropy-based Fuzzy Clustering (EFC) [3] and so on. On the other hand, 'hard' clusters possess well-defined boundary, which is seen in K-means (KM) [4], Hierarchical methods [5] and so forth. Choosing correct algorithm has always been a research challenge [6]. In this paper, we compare the performance of one 'soft' (e.g., FCM) and one 'hard' clustering i.e., KM technique on three standard datasets of various sizes. These datasets are *Iris* (150 x 4), *Wine* (178 x 13) and *Lens* (24 x 4), obtained from UCI machine learning database [7]. Performances of FCM and KM are also compared based on Manhattan (MH) and Euclidean (ED) measures.

The objective of this study is to examine the best *'clustering methods-distance measure'* combinations in terms of (a) 'quantitative clustering' (which are checked with Sensitivity, Specificity, Precision and Accuracy measures); (b) 'speed' (examined by measuring the run time); and (c) 'quality' of the cluster in terms of compactness and distinctiveness (i.e., how far one cluster is situated from another cluster).

2 Methodology

The objective of this study is to compare the performance of FCM and KM on three standard datasets, such as Iris, Wine and Lens. In order to accomplish the task, the algorithms are developed in 'C' language and implemented.

Working Principle of FCM Algorithm:

1. *Initialize* $U=[u_{ij}]$ *matrix,* $U^{(0)}$
2. *At k-step: calculate the centroid* $C^{(k)}=[c_j]$ *with* $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} .x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \tag{1}$$

3. *Update* $U^{(k)}$, $U^{(k+1)}$

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}} \tag{2}$$

4. *If* $\| U^{(k+1)} - U^{(k)} \| < \theta$ *then STOP; otherwise return to step 2.Here, 'm' is the fuzziness parameter.*

Working Principle of KM Algorithm:

For 'M' sample vectors $\{x_1,x_2,\ldots,x_M\}$ falling into 'k' compact clusters (k<M)

Let 'm_i' be the mean of the vectors in cluster 'i'

If $\|x-m_i\|$ is the minimum of all 'k' distances

insert 'x' into the respective cluster

until there is no change in any 'm'.

To note how the distance measures influence the clustering tasks, two distance measures have been used, e.g., Manhattan (MH) and Euclidean (ED). These distances follow L^P -norm (see equation 1).

$$\|x\|_p = \left[\sum_{i=1}^{k}|x_i|^p\right]^{\frac{1}{p}} \tag{3}$$

In this equation, 'x_i' are the number of data points. Now for 'p'=1 we get MH and for 'p'=2 it is ED.

The performance of these two techniques has been compared based on the following parameters:

1. Sensitivity: $\frac{T_P}{P}$ (4)

In this equation T_P is 'true +' and 'P' is '+'.

2. Specificity: $\frac{T_N}{N}$ (5)

In this equation T_N is 'true -' and 'N' is '-'.

3. Precision: $\frac{T_P}{T_P+F_P}$ (6)

In this equation Fp is 'false +'

4. Accuracy: $\frac{n}{n*}$ (7)

Here '*n*' denotes the total number of correctly classified data and '*n**' is the total number of data.

5. Run time: the amount of time (in seconds) spent to run the algorithm in a PIV (core2duo) PC.

6. Intra cluster distance: $\frac{1}{n*}\sqrt{\sum_{i=1}^{n*}\|x_i-c\|^2}$ (8)

In this equation, 'c' denotes the cluster center (or, centroid). From this equation, we can infer if such distance is low, the respective clusters are more 'compact'.

7. Inter cluster distance: $d = \left[\left\| c_{ij} - c_{ji} \right\|^{p}\right]^{\frac{1}{p}}$ (9) $(i \neq j); p = 2$

   Here, 'm' is the desired number of clusters. The desired inter cluster distance should be high to infer that the clusters are not overlapped.

3 Results and Discussions

In this section, the experimental results are displayed and explained as follows. Table 1-3 shows the performance results on three datasets.

Table 1. Performance analysis of FCM and KM clustering techniques on **Iris** data

		FCM		**KM**	
Parameters	**Cluster-info**	**MH**	**ED**	**MH**	**ED**
Sensitivity	CL1	1	1	1	1
	CL2	0.94	0.94	0.84	0.84
	CL3	0.72	0.72	0.84	0.84
	Average	0.8866	0.8866	**0.8933**	**0.8933**
Specificity	CL1	1	1	1	1
	CL2	0.86	0.86	0.82	0.82
	CL3	0.97	0.97	0.9	0.9
	Average	**0.9433**	**0.9433**	0.9066	0.9066
Precision	CL1	1	1	1	1
	CL2	0.7705	0.7705	0.84	0.84
	CL3	0.9231	0.9231	0.84	0.84
	Average	**0.8978**	**0.8978**	0.8933	0.8933
Accuracy	*Average*	0.8867	0.8867	**0.8933**	**0.8933**
Time	*Average*	0.2321	0.2321	**0.1281**	**0.1281**
Intra cluster Distance	CL1	0.895	0.487	1.261	**0.696**
	CL2	1.261	0.695	0.908	**0.493**
	CL3	1.184	0.65	1.101	**0.605**
Inter cluster Distance	CL1-CL2	**1.487**	0.775	1.47	0.767
	CL2-CL3	**2.129**	1.174	0.881	0.448
	CL1-CL3	**0.951**	0.487	2.012	1.107

Table 2. Performance analysis of FCM and KM clustering techniques on **Wine** data

		FCM		**KM**	
Parameters	**Cluster-info**	**MH**	**ED**	**MH**	**ED**
Sensitivity	CL1	1	1	1	1
	CL2	0.9014	0.9014	0.9155	0.9155
	CL3	1	1	1	1
	Average	0.9671	0.9671	**0.9718**	**0.9718**
Specificity	CL1	0.9748	0.9748	0.9748	0.9748
	CL2	1	1	1	1
	CL3	0.9692	0.9692	0.9769	0.9769
	Average	0.9813	0.9813	**0.9839**	**0.9839**
Precision	CL1	0.9516	0.9516	0.9516	0.9516
	CL2	1	1	1	1
	CL3	0.9231	0.9231	0.9412	0.9412
	Average	0.9582	0.9582	**0.9642**	**0.9642**
Accuracy	*Average*	0.9607	0.9607	**0.9663**	**0.9663**
Time	*Average*	0.5007	0.5007	**0.1976**	**0.1976**
Inter cluster Distance	CL1	**2.28**	0.772	2.2109	0.7430
	CL2	**3.523**	1.105	2.6007	0.8259
	CL3	**2.701**	0.85	3.4510	1.0867
Intra cluster Distance	CL1-CL2	**2.29**	0.786	2.5148	0.8338
	CL2-CL3	**2.564**	0.849	2.2870	0.7825
	CL1-CL3	**2.797**	0.849	2.7950	0.9166

In Iris data, KM clusters with greater accuracy than FCM, which renders negligibly better precision. While testing the run time, it may note that KM takes much less time compared to FCM for both the MH and ED distances. FCM-ED is able to produce most compact clusters, while FCM-MH yields most distinct clusters. In Wine data, again the first five parameters (sensitivity, specificity, precision, accuracy, and run time) show similar results as seen in Iris data. KM-ED combination is able to produce the most compact clusters. Both KM-MH and FCM-MH is able to produce most distinct clusters. Similar results (as seen in Wine) could be seen in Lens data as well. KM-ED combination is able to produce most compact clusters, while FCM-MH produces most distinct clusters.

Figures 1(a) and (b) show the FCM and KM-based classification plots of MH and ED distance measures on Iris datasets. Similarly, classification plots have been obtained with Wine and Lens datasets (shown in fig. 2 and 3, respectively).

Table 3. Performance analysis of FCM and KM clustering techniques on **Lens** data

		FCM		KM	
Parameters	**Cluster-info**	**MH**	**ED**	**MH**	**ED**
Sensitivity	CL1	0.5	0.5	0.75	0.75
	CL2	1	1	1	1
	CL3	1	1	1	1
	Average	0.8333	0.8333	**0.9166**	**0.9166**
Specificity	CL1	1	1	0.9748	0.9748
	CL2	0.75	0.75	0.875	0.875
	CL3	0.75	0.75	0.875	0.875
	Average	0.8333	0.8333	**0.9082**	**0.9082**
Precision	CL1	0.33	0.33	0.6	0.6
	CL2	1	1	1	1
	CL3	1	1	1	1
	Average	0.7766	0.7766	**0.8666**	**0.8666**
Accuracy	*Average*	0.833	0.833	**0.917**	**0.917**
Time	*Average*	0.0429	0.0429	**0.0296**	**0.0296**
Inter cluster Distance	CL1	1.4889	1.3609	2.2333	**1.2871**
	CL2	1.6174	1.0599	1.7375	**0.9521**
	CL3	2.5152	1.3514	1.8958	**1.1197**
Intra cluster Distance	CL1-CL2	**4.7358**	3.0057	4.9583	3.0016
	CL2-CL3	**4.3081**	2.5404	5.2583	2.8585
	CL1-CL3	**5.1177**	2.8430	3.8958	2.3852

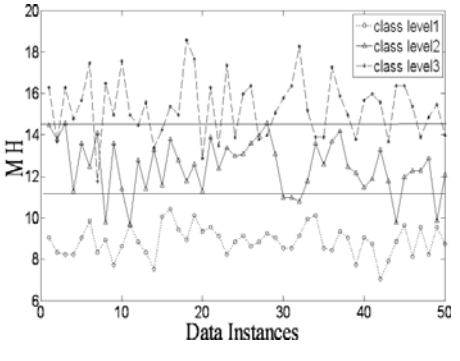


Fig. 1a. Classification of **Iris** data using **FCM** algorithm and **MH** distance

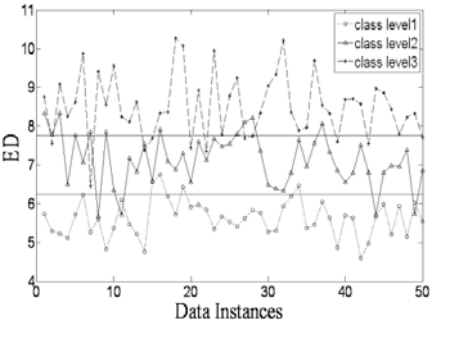


Fig. 1b. Classification of **Iris** data using **FCM** algorithm and **ED** distance

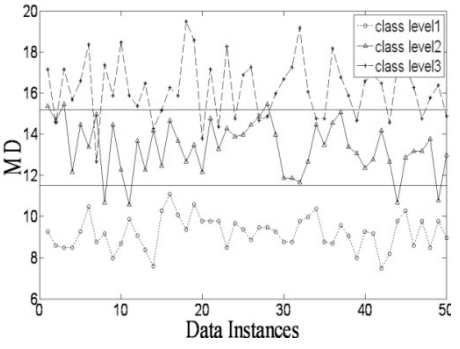


Fig. 2a. Classification of **Iris** data using **KM** algorithm and **MH** distance

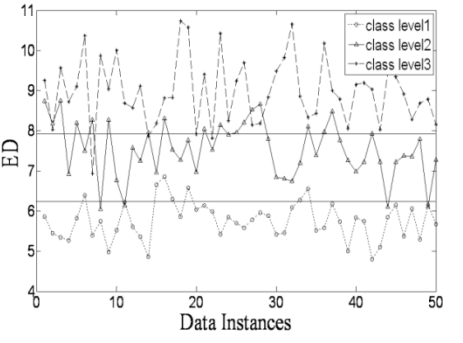


Fig. 2b. Classification of **Iris** data using **KM** algorithm and **ED** distance

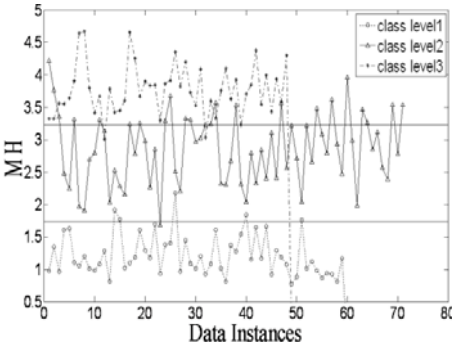


Fig. 3a. Classification of **Wine** data using **FCM** algorithm and **MH** distance

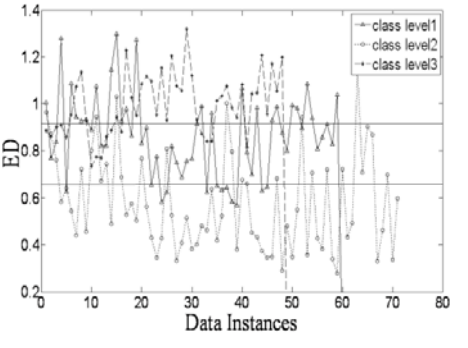


Fig. 3b. Classification of **Wine** data using **FCM** algorithm and **ED** distance

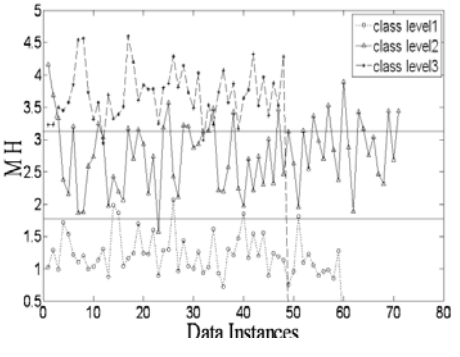


Fig. 4a. Classification of **Wine** data using **KM** algorithm and **MH** distance

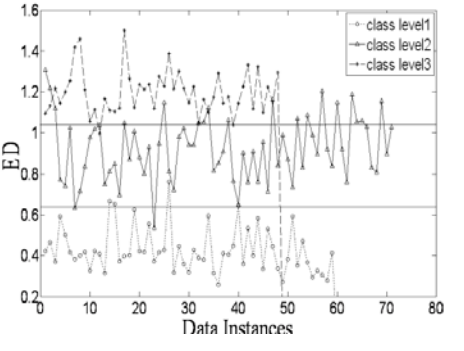


Fig. 4b. Classification of **Wine** data using **KM** algorithm and **ED** distance

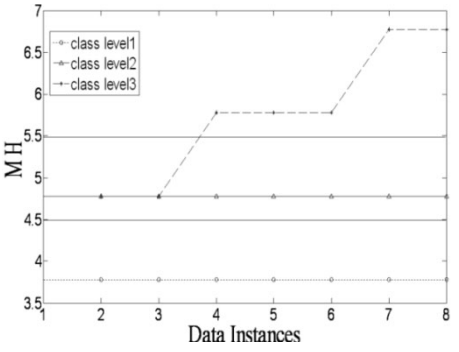


Fig. 5a. Classification of **Lens** data using **FCM** algorithm and **MH** distance

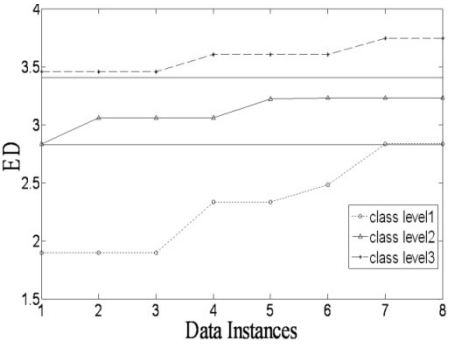


Fig. 5b. Classification of **Lens** data using **FCM** algorithm and **ED** distance

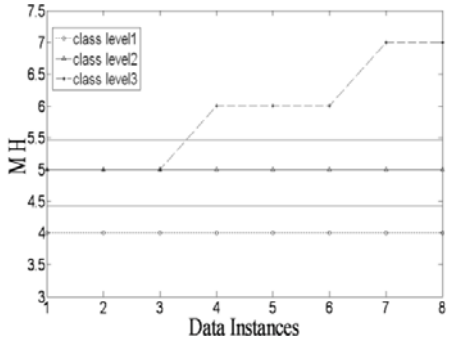


Fig. 6a. Classification of **Lens** data using **KM** algorithm and **MH** distance

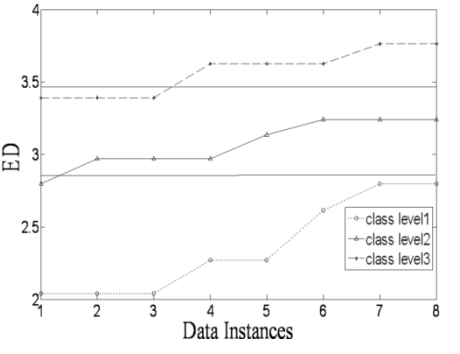


Fig. 6b. Classification of **Lens** data using **KM** algorithm and **ED** distance

Computational time has finally been computed. Figure 7(a), (b), and (c) shows the respective plots for MH distance as it is obvious that the amount of computation in MH is much less that the ED.

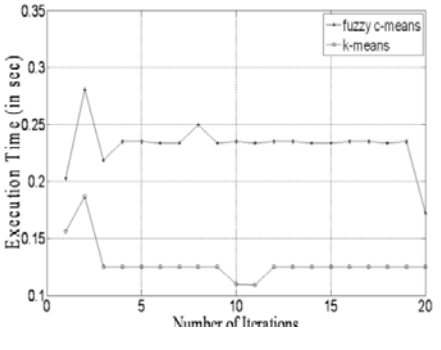


Fig. 7a. Run time plots on **Iris** data

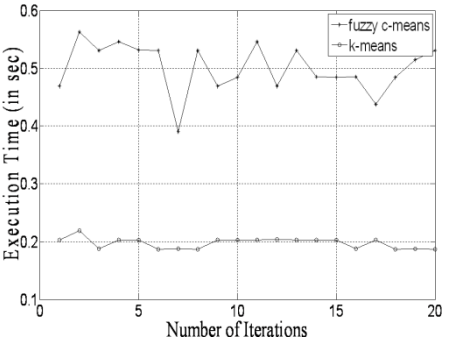


Fig. 7b. Run time plots on **Wine** data

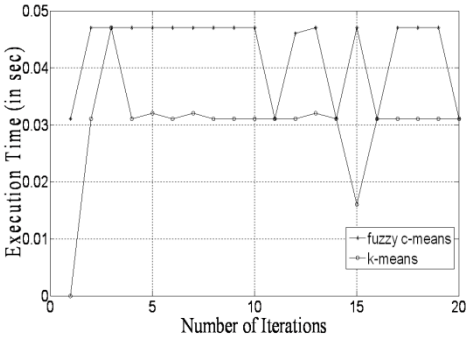


Fig. 7c. Run time plots on **Lens** data

4 Conclusions and Future Work

The results reveal that on Iris data higher precision values for clustering have been obtained with both MH and ED. With FCM-MH combination, more distinct clusters are produced. On the other hand, higher accuracy could be revealed with KM-MH and KM-ED combinations. With the same combination, KM is able to produce high quality clusters with minimum run time. Finally, KM-ED combination is able to yield most compact clusters. In the Wine data, with FCM-MH combination produces most compact and distinct clusters with greater accuracy and minimum time. In the Lens data, FCM-MH can produce the most distinct clusters. For the other parameters, KM performs better than the FCM algorithm. From the results, it is observed that, overall, KM outperforms FCM.

In future, the algorithms could be implemented on real-life clinical data, which are much subjective in nature. Therefore, it would be challenging to choose the right clustering-distance measure approach. Currently the authors are working on this topic.

References

[1] Bezdek, J.C.: Fuzzy mathematics in pattern classification. Applied Mathematics Centre, Cornell University, Ithaca. PhD thesis (1973)

[2] Keller, J., Gary, M.R., Givens, J.A.: A fuzzy k-nearest neighbor algorithm. IEEE Tr. Syst. Man Cyber. 15(4), 580–585 (1985)

[3] Yao, J., Dash, M., Tan, S.T., Liu, H.: Entropy-based fuzzy clustering and fuzzy modeling. Fuzzy Sets and Systems 113, 381–388 (2000)

[4] MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, CA, pp. 281–297 (1967)

[5] Sibson, R.: SLINK: an optimally efficient algorithm for the single-link cluster method. The Computer Journal (British Computer Society) 16(1), 364–366 (1973)

[6] Chattopadhyay, S., Pratihar, D.K., De Sarkar, S.C.: A comparative study of fuzzy C-means algorithm and entropy-based fuzzy clustering algorithm. Computing and Informatics 30(4), 701–720 (2011)
[7] http://archive.ics.uci.edu/ml/ (Online; last accessed on December 23, 2011)
[8] Han, J., Kamber, M. (eds.): Data Mining Concepts and Techniques, 2nd edn. Elsevier, San Fransisco (2006)

Comparative Analysis of Diverse Approaches for Air Target Classification Based on Radar Track Data

Manish Garg and Upasna Singh

Department of Computer Engineering
Defence Institute of Advanced Technology, Pune
{cse10manish,upasnasingh}@diat.ac.in

Abstract. Air Target Classification in a hostile scenario will be a decisive factor for threat evaluation and weapon assignment. Stealth technology denies any high frequency based regime for such classification. It is observed that kinematics of an air target is one thing that cannot be deceived. The present study makes an attempt to ascertain an appropriate Classification algorithm. On the basis of certain significant feature vectors the classifier classifies the data set of an air target into a target class. Feature vectors are derived from the Radar Track Data using Matlab code. The work presented here aims to compare the predictability importance of features using different classification algorithms.

Keywords: Air Target Classification, Feature Vectors, Kinematics, Predictability Importance, Radar Track data.

1 Introduction

The problem of classification and identification of aircraft which do not broadcast their own identity is of recurring interest. The capability to identify an air target as a fighter, a transport aircraft, a rotary wing aircraft or an unmanned aerial vehicle (UAV) is known as the air target classification. This capability acquires enormous importance when the air target is not cooperating and does not want to share its identity. At this moment it is of interest to all the people in the field of Air Defence to classify the air target and deal with the situation efficiently. For non coop-erative target recognition, considerable effort has been expended. Unfortunately, no accurate and practicable technique has been developed which is appropriate to small "real-time" systems ([1]-[9]).

In [12] various approaches for air target classification are analysed- Use of polar metric data, one dimensional High Resolution Range Profiles (HRRPs) or two dimensional Inverse Synthetic Aperture Radar (ISAR). Most of the work has been undertaken in the high frequency regime and involves extraction of scattering centers for identification of targets. It was inferred that the kinematics of an air target is a dependable source for air target classification.

In this paper, we quantify kinematic information from radar track data, as received from a radar tracker, and use it to assess the likelihood of a tracked target. Aircraft have unique physical attributes that characterize their angular and translational motion

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 461–470.
springerlink.com

due to applied input forces. The analysis of these unique significant features leads to the classification.

The paper is organized in terms of six sections. It starts with the introduction in section 1 followed by the basic constructs on section 2. In section 3 we have explained the methodology which uses several feature extraction parameters for classifying air targets. Thereafter experimental evaluation with different classification algorithms are shown in section 4. The comparative analysis in section 5 is the overall result of the experiments performed. Section 6 concludes the work done.

2 Basic Constructs

Radar Track Data: A radar tracker is a component of a radar system, or an associated command and control (C2) system that associates consecutive radar observations of the same target into tracks. The sequence of data, as received from a radar tracker, is known as Radar track Data.

3 Methodology

Different classes of aircraft (e.g. fighter plane, commercial passenger aircraft) differ in their geometry, size, and flight envelope and in particular in their maneuvering capabilities [11]. The acceleration capabilities of various targets can most naturally be incorporated into the target's discrete-time dynamical state equation as input terms [8], while the rotational (angular dynamical) properties of aircraft can be incorporated as extra states [9]. [12] proposes methodology as shown below:

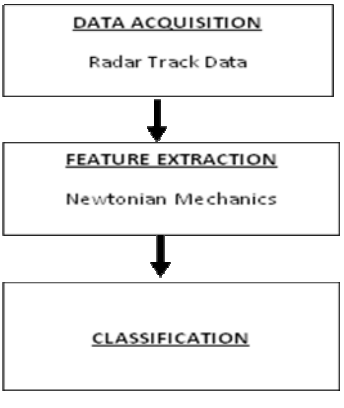


Fig. 1. Proposed Methodology

Data Acquisition. The dynamical behavior of the target is embedded in the ob-served radar data and extraction of this information is a basis for target classification. The data received is a sequence of plots made in space. It may be acquired in any form of coordinates, say polar metric, Earth Centered Earth Fixed (ecef), Cartesian etc.

Feature Extraction. An aircraft in motion can be considered to have six degrees of freedom (Three axes x-y-z for translational move and moment about these three axes for rotational move). Sensor data are discrete time measurements thus motivating a discrete time approximation of the linear equations of motion. Thus, we consider the aircraft kinematic model as described by the linear discrete time system

$$X_{(k+1)} = \text{fn}\,(X_k) \tag{1}$$

where the state X is a 4 dimensional vector i.e. (x,y,z,t) . Using Newtonian mechanics, the translational movement is analyzed in terms of rate of change of space wrt time.

$$[V_x,V_y,V_z] = d/dt\,[x,y,z] \tag{2}$$

$$V= (V_x+V_y+V_z)^{1/2} \tag{3}$$

where V denotes the velocity and the subscript denotes the component along a axes.This would give the relative velocity of the object about the three axes. Further analysis in terms of rate of change of velocity wrt time calculates the acceleration along the 3-axes and at the same time illustrates the maneuver of the target.

$$[A_x,A_y,A_z] = d/dt\,(V_x,V_y,V_z) \tag{4}$$

$$A= (A_x+A_y+A_z)^{1/2} \tag{5}$$

where A denotes the acceleration and the subscript denotes the component along a axes.Further examination of the three dimensional trajectory of the target, its Curvature at any instance is calculated. For a parametrically defined space curve in three-dimensions given in Cartesian coordinates (x(t),y(t),z(t)), the curvature is

$$\kappa = \frac{\sqrt{(z''y' - y''z')^2 + (x''z' - z''x')^2 + (y''x' - x''y')^2}}{(x'^2 + y'^2 + z'^2)^{3/2}}. \tag{6}$$

where the prime denotes differentiation with respect to time t. [10] has brought to light the maneuvering capability of a air target in terms of its capability to accelerate. To examine the relationship between the velocity at which an air target can make a turn, the centrifugal acceleration ϵ is calculated

$$\epsilon = V^2/\kappa \tag{7}$$

Eventually, the set of feature vectors that is contemplated to be significant to classify the air target is given as follows:

$$\varepsilon = [z, V_x,V_y,V_z, A_x,A_y,A_z,V,A, \kappa, \epsilon] \tag{8}$$

Thus the flight state of an air target depicted as ε can be seen as a function of its altitude, velocity and maneuvering capability. It is also seen in Figure 2 that although the flight envelop of an aircraft is bound by its features, there are flight envelops overlapping.

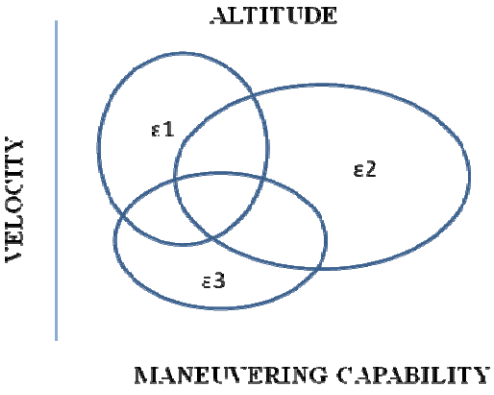


Fig. 2. Illustration of flight envelopes of targets belonging to different classes

Classification. Further step in the target identification process concerns the transformation of a set of entity attributes into a label describing the target identity. Davis,s law states that ' For every tool there is a task perfectly suited to it (Davis and King,1977).The purpose of comparison of classification algorithm is to determine the apt algorithm in a coned environment where an air target is trying to deceive and deliberately flies in the overlapping flight envelop zone.

4 Experimental Analysis

Similar to [12] four target classes are identified for classification:

1. Unmanned Aerial Vehicle.
2. Rotor Wing Aircraft.
3. Transport Aircraft.
4. Fighter Aircraft.

Data Sets Used: Four Samples of each class are used. Each sample has four attributes: [time,x,y,z] i.e. Radar Track Data which is shown in Table 4.1 gives the instantaneous location of air target in space. Each sample has more than 1000 records. Here the [x,y,z] coordi-nates are given in Earth-Centered, Earth-Fixed (ecef) format.

Table 1. Sample Radar Track Data

Time	X	y	Z
0	1324333	5395816	3123928
1	1324239	5395846	3123917
…	……	……	……
1005	1323862	5395966	3123871

Using Matlab code, the RTD is converted to feature vectors as discussed in (8). Here coordinate reference system used is ENU (East, North, Up). Similarly, all the samples are converted to feature vectors as shown.

Table 1. Sample feature vectors as derived from Radar Track Data

Altitude	Vx	Vy	Vz	Ax	Ay	Az	V	A	€
650.2329	98.4140	-13.0771	0.818962	0	-0.87184	-0.4897	99.2824	1	0.993843
……..	……	……	……	……	……	……..	……..	…..	…….
……..	……	……	……	……	……	……..	……..	…..	…….
……..	……	……	……	……	……	……..	……..	…	…….
652.4208	98.4140	-13.0771	0.818962	3.64E-12	-0.87184	-0.48979	99.2824	1	0.993843

[12] considered reducing features to disallow deception to the air target. It may not be prudent to disregard velocity and altitude to deny deception to the target. However we considered that reducing the predictability importance of any one feature will be more effective. Input to all algorithm models is a mixed sample of all four classes containing all 9 input attributes and 10th target field class. For classification, we have used IBM SPSS Modeler version 14.1. Here are the results of certain algorithms considered for classification:

Algorithm 1 (CRT). It is seen in Figure 3 that velocity and altitude are the most important field.

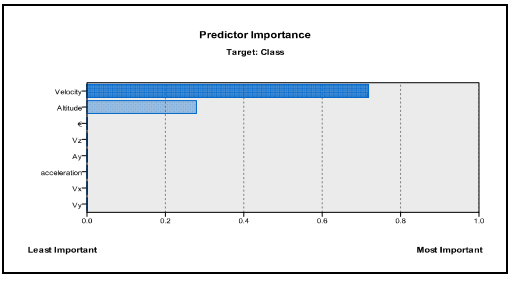


Fig. 3. Predictor analysis showing inter se importance of the feature vectors for CRT model

Comparing $R-Class with Class

Correct	14,811	100%
Wrong	0	0%
Total	14,811	

Algorithm 2 (QUEST). Now it is seen in Figure 4 that velocity and altitude are still the most important feature.

Comparing $R-Class with Class

Correct	14,810	99.99%
Wrong	1	0.01%
Total	14,811	

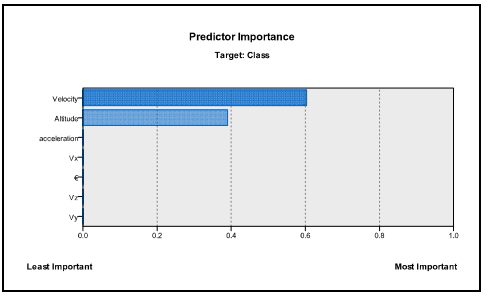


Fig. 4. Predictor analysis showing inter se importance of the feature vectors for QUEST model.

Algorithm 3 (CHAID). It is seen in Figure 5 that velocity is the most important field used for classification.

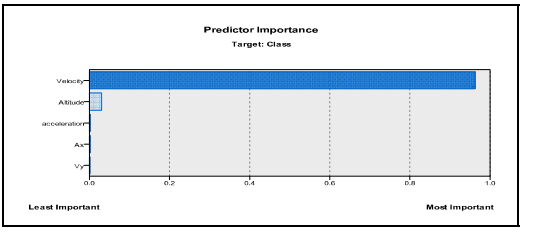


Fig. 5. Predictor analysis showing inter se importance of the feature vectors for CHAID model.

In addition to velocity using very little assistance from altitude, model gives following results for output field Class

Comparing $R-Class with Class

Correct	14,572	98.39%
Wrong	239	1.61%
Total	14,811	

Algorithm 4 (Artificial Neural Network). Now it is seen in Figure 6 that predictor importance is equally distributed for classification.

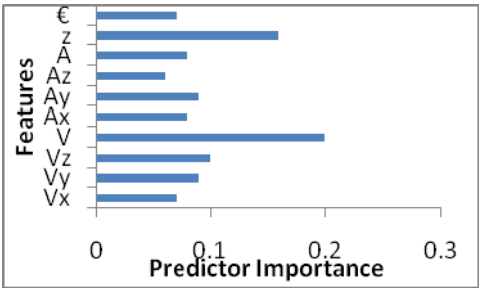


Fig. 6. Predictor analysis showing inter se importance of the feature vectors for ANN model

Results for output field Class
Comparing $N-Class with Class

Correct	14,775	99.76%
Wrong	36	0.24%
Total	14,811	

Algorithm 5 (C 5.0). It is seen in Figure 7 that velocity and altitude are the most important field.

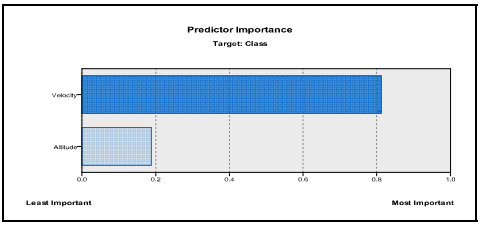


Fig. 7. Predictor analysis showing inter se importance of the feature vectors for C 5.0 model

Results for output field Class
Comparing $C-Class with Class

Correct	14,809	99.99%
Wrong	2	0.01%
Total	14,811	

Algorithm 6 (Discriminant). Now it is seen in Figure 8 that predictor importance is maximum for velocity and altitude for classification.
Results for output field Class
Comparing $D-Class with Class

Correct	14,551	98.24%
Wrong	260	1.76%
Total	14,811	

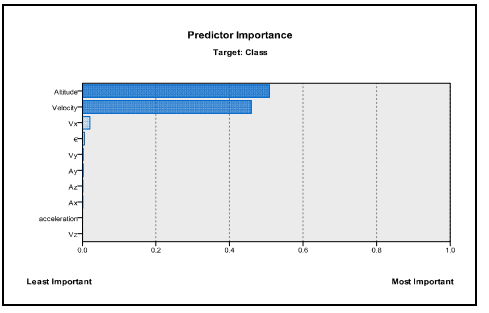


Fig. 8. Predictor analysis showing inter se importance for Discriminant model

Algorithm 7 (Bayes Network). Now it is seen in Figure 9 that predictor importance is equally distributed for classification.

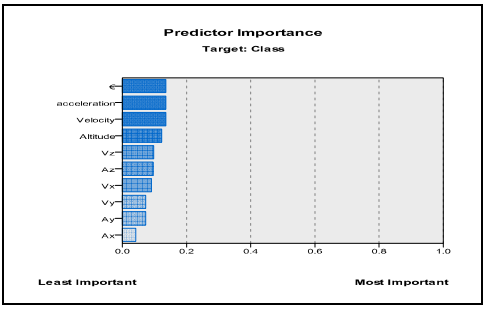


Fig. 9. Predictor analysis showing inter se importance for Bayesian Network model

Results for output field Class

Comparing $B-Class with Class

Correct	13,712	92.58%
Wrong	1,099	7.42%
Total	14,811	

5 Results

It is observed that Decision tree based classification algorithms are classifying based on only two features' attributes, mainly velocity and altitude. In case of hostile environment, where an air target tries to deceive the tracking and classification system, this form of classification based on decision tree may give the wrong results. To handle the deception, relying on any one or two features is not the right thing to do. It implies that decision tree is not an appropriate algorithm for air target classification. On the contrary, Artificial Neural Network and Bayes Network give adequate weightage to all feature vectors. Therefore, it can be deduced that a neural or Bayes network can reliably deny deception to an air target trying to con.

Table 3. Comparison of results showing various input fields and their respective classification correctness

Classification Algorithm	Result		*Inter se* predictability Importance									
	Correct(%)	Wrong(%)	V_x	V_y	V_z	V	A_x	A_y	A_z	A	z	ϵ
CRT	100	0	-	-	-	0.72	-	-	-	-	0.28	-
QUEST	99.99	0.01	-	-	-	0.6	-	-	-	-	0.39	-
CHAID	98.39	1.61	-	-	-	0.96	-	-	-	-	0.03	-
ANN	99.76	0.24	0.07	0.09	0.1	0.2	0.1	0.09	0.06	0.08	0.16	0.07
C 5.0	99.99	0.01	-	-	-	0.81	-	-	-	-	0.19	-
Discriminant	98.24	1.76	0.02	-	-	0.46	-	-	-	-	0.51	0.01
Bayes Network	92.58	7.42	0.09	0.07	0.1	0.13	0	0.07	0.1	0.14	0.12	0.14

6 Conclusion

The information from the radar, i.e. its location in space at an instant, can provide indications of the target type. It is observed that the flight state i.e. its altitude, velocity and acceleration at an instant is strictly bound by a flight envelop. Although these envelops are overlapping in limited dimension scenario there are certain peculiar characteristics, of the target class, that are able to classify air target. This study attempts to identify appropriate classification algorithm that would be able to perform in an environment where an air target is trying to replicate the flying characteristics of another class.

References

[1] Borden, B.H.: Enhanced Range Profiles for Radar-Based Target Classification using Monopulse Tracking Statistics. IEEE Transactions on Antennas and Propagation 43(8) (August 1995)

[2] Hu, R., Zhu, Z.: Researches on Radar Target Classification Based on High Resolution Range Profiles. In: Proc. of the IEEE Aerospace Conference, vol. 4, pp. 1243–1251 (2002)

[3] Lanterman, A.D.: Tracking and Recognition of Airborne Targets via Commercial television and FM radio signals. In: Proc. of SPIE Acquisition, Tracking, and Pointing, vol. 3692, pp. 189–198 (1999)

[4] Herman, S., Moulin, P.: A Particle Filtering Approach to FM-Band Passive Radar Tracking and Automatic Target Recognition. In: Proc. of the IEEE Aerospace Conference, vol. 4, pp. 1789–1808 (2002)

[5] Cutaia, N.J., O'Sullivan, J.A.: Automatic Target Recognition using Kinematic Priors. IEEE Transactions on Aerospace and Electronics Engineering, AES-23 (May 1987)

[6] Edwards, G., Tate, J.P.: Target Recognition and Classification using Neural Networks. IEEE Transactions on Antennas and Propagation 43(8), 1439–1442 (2002)
[7] Kouemou, G.: Radar Target Classification Technologies. In: INTECH, Croatia, downloaded from SCIYO.COM, p. 410 (December 2009) ISBN 978-953-307-029-2
[8] Bogler, P.L.: Tracking a maneuvering Target using input estimation. IEEE Transactions on Aerospace and Electronics Engineering, AES-23, pp. 298–310 (May 1987)
[9] Cutaia, N.J., O'Sullivan, J.A.: Identification of maneuvering aircraft using class de-pendent kinematic model. In: Research Monograph, ESSRL-95-13, Electronic Signals and Systems Research Laboratory, Department of Electrical Eng., Washington University, St. Louis, MO (May 1995)
[10] Whitford, R.: Design for Air Combat. Janes (1987)
[11] Stinton, D.: The Anatomy of the Aeroplane. BSP (1985)
[12] Garg, M., Singh, U.: C & R Tree based Air Target Classification using Kinematics. In: National Conference on Research Trends in Computer Science and Technology (NCRTCST), IJCCT_Vol3Iss1/IJCCT_Paper_3 (2012)

Speed Optimization in an Unplanned Lane Traffic Using Swarm Intelligence and Population Knowledge Base Oriented Performance Analysis

Prasun Ghosal[1], Arijit Chakraborty[2], and Sabyasachee Banerjee[2]

[1] Department of Information Technology, Bengal Engineering and Science University, Shibpur, Howrah 711110, WB, India
[2] Department of Computer Science and Engineering, Heritage Institute of Technology, Kolkata, WB, India
{prasung,arijitchakraborty.besu,sabyasachee.banerjee}@gmail.com

Abstract. Comparative Analysis of Speed Optimization Technique in Unplanned Traffic is a very promising research problem. Searching for an efficient optimization method to increase the degree of speed optimization and thereby increasing the traffic flow in an unplanned zone is a widely concerning issue. However, there has been a limited research effort on the optimization of the lane usage with speed optimization. This paper presents a novel technique to solve the problem optimally using the knowledge base analysis of speeds of vehicles, using partial modification of Swarm Intelligence which, in turn will act as a guide for design of lanes optimally to provide better optimized traffic with less number of transitions between lanes.

1 Introduction

The challenges of the accidental and congested lane design system are to move traffic safely and efficiently, although, highways and motor vehicles are designed to operate safely at speed. The purpose of our investigation is to create predictive models for different types of speed optimization techniques on lane, based on infrastructural design and traffic intensity. In this paper, the results for all transition points and vehicle's lane transition for speed optimization is discussed.

The Analysis starts with identifying main issues and element of the problem in hand which are as follows.

- Entry zones,
- Transition points, and
- Exit zones.

Most of the traditional approach for tackling the problem in hand is based on deterministic models which can be efficient and more or less accurate at times, but to achieve optimality of solution deterministically, at all time, seems to be far from reality till now.

Apart from that, making the lanes at their optimal average speed at any point of time using previous knowledge and current information is a major highlight presented in this paper.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 471–480.
springerlink.com

The overall organization of the paper is as follows. Section 2 discusses about the background of the present work with a description of the related works done so far in this area, and pointing out the drawbacks of the existing solutions. In the next section, the problem formulation and proposed algorithms are represented. First algorithm does not consider the concept of population knowledge base, and second one with the population knowledge base. Experimental results and observations are represented in section 4. Finally, section 5 concludes the paper with possible future directions of work.

2 Background and Motivation

2.1 Related Works

The paper proposed by Jake Kononov, Barbara Bailey, and Bryan K. Allery, first explores the relationship between safety and congestion and then examines the relationship between safety and the number of lanes on urban freeways.

The relationship between safety and congestion on urban freeways was explored with the use of safety performance functions [SPF] calibrated for multilane freeways in Colorado, California, Texas.

The Focus of most SPF modeling efforts to date has been on the statistical technique and the underlying probability distributions. The modeling process was informed by the consideration of the traffic operations parameters described by the Highway Capacity Manual. [1]

H Ludvigsen, Danish Road Directorate, DK; J Mertner, COWI A/S, DK, 2006, published, Differentiated speed limits allowing higher speed at certain road sections whilst maintaining the safety standards are presently being applied in Denmark.

The typical odds that higher speed limits will increase the number of accidents must thus be beaten by the project.

That paper presented the methodology and findings of a project carried out by the Danish Road Directorate and COWI aimed at identifying potential sections where the speed limit could be increased from 80 km/h to 90 km/h without jeopardizing road safety and where only minor and cheaper measures are necessary. Thus it described how to systematically assess the road network when the speed limit is to be increased. [2].

C.J. Messer and D.B. Fambro, 1977, presented a new critical lane analysis as a guide for designing signalized intersections to serve rush-hour traffic demands.

Physical design and signalization alternatives are identified, and methods for evaluation are provided. The procedures used to convert traffic volume data for the design year into equivalent turning movement volumes are described, and all volumes are then converted into equivalent through-automobile volumes.

The critical lane analysis technique is applied to the proposed design and signalization plan. The resulting sum of critical lane volumes is then checked against established maximum values for each level of service (A, B, C, D, E) to determine the acceptability of the design. [3].

2.2 Drawbacks of Existing Solutions

Many traditional speed-optimizing algorithms for lanes were proposed earlier to optimize deterministic problems. But these algorithms didn't show their ability to use their

previous knowledge to tackle the inherent randomness in the traffic systems. Therefore, to handle with such random realistic situation and generate some efficient solution, good computational models of the same problem as well as good heuristics are required.

This article is divided into two major sections: -

In first part, simulation algorithm will provide us with no. Of lanes required moving the traffic at optimal speed in each proposed lane.

Second part, deals with knowledge obtained from the frist part to make the lane transitions less in number making it nearer towards the desired goal.

3 Problem Formulations and Proposed Algorithms

3.1 Problem Description

Description of Figure 1:

Figure 1**, T**hree vertical lanes that are unidirectional, and A = {a1, a2… an}, B = {b1, b2 ….bn}, C = {c1, c2,….,cn}, three lanes. I, II, III are the transition points through which vehicles can overtake its preceding vehicle with lesser speed and then immediately moves to its original lane. i.e. I from lane A to B or B to A and II, III are from B to C or C to B. Here we assume that each and every lane's car speed is greater than 0 kmph. If speed of any car is less than or equal to 0 kmph then we assume that there may be problem.

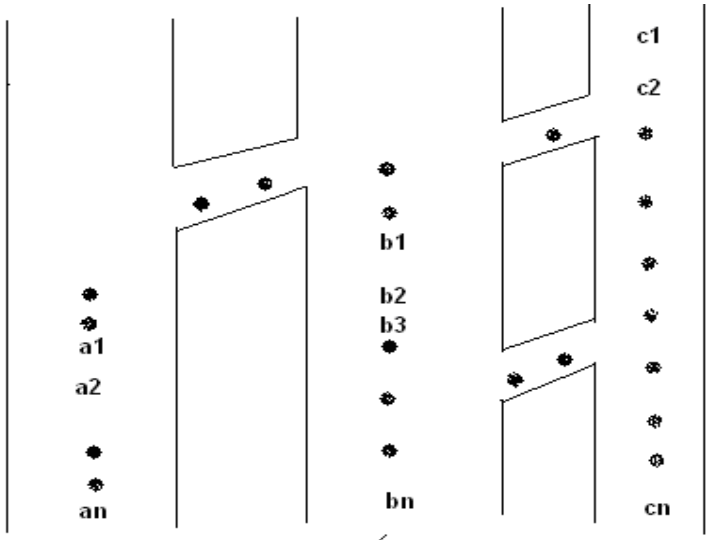


Fig. 1. Vertical lanes are unidirectional and with the property of the three lanes with transition points

3.2 Problem Formulation

Random movement of vehicle in rush hour traffic are required to be frame up in optimal no. lanes with respect to number of transitions between lanes so that each lane have optimal speed.

Bio inspired algorithms like swarm intelligence (.i.e. Ant Colony Optimization) technique used here with speed of the 'vehicle' acting as a pheromone to solve the problem in hand.

To maintain the optimality of a solution in a heuristic search using population information as a knowledge base is used in the proposed algorithms.

3.3 Proposed Algorithms

3.3.1 Algorithm I

Initial Assumptions

- There will be no change in the speed of the vehicle
- In case of sudden change of speed, accommodate the speed of previous slower vehicle.
- Any vehicle having speed equivalent of 0 is discarded from the initial sample or population

Details of the Proposed Algorithm

The major keynotes and functionality of the proposed algorithm are as follows: -

Step 1 is taking input from sensors, like the current speed of the vehicle, arrival time etc., and, counting the number vehicles the user has entered.

Step 2 is categorizing the vehicles depending on their current speed.

Step 3 is checking total how many numbers of lanes will be required for our sample data in an unplanned zone, and, which vehicle is moving in which lane.

Step 4 is checking total number of transitions i.e. at which point of the lane and from which lane to where the transition will occur.

Table 1. Symbolic Interpretation used in algorithms

Symbols used	Meaning
V_i	Velocity of vehicle i
V_j	Velocity of vehicle j
L_i	Lane of the vehicle i
L_j	Lane of the vehicle j
L_1	Lane of the 1st vehicle.
type(i)	Category of Vehicle i
t	Arrival time difference between a high and low speed vehicles
t1	Time interval to overtake a vehicles at lower speed
d	Distance covered by low speed Vehicle
d1	Distance covered by high speeding Vehicle
B_n	Buffer of Lane n
Count	Total no. Vehicle in unplanned traffic
Count1	Total no. Lanes for optimal speed
Count2	Total number Of transition

Pseudo Code (Algorithm I)

Input: Details of vehicles, Current speed of the vehicle, arrival time.
Output: Category of the vehicle, Number of lanes will be required, Number of transitions.

Step 1.1: Set count = 1; /*Used to count the number of vehicles. */
Step 1.2: get_ input (); /*Enter Details of vehicles, current speed, arrival time and store it into a record. */
Step 1.3: Continue Step 1.1 until sensor stops to give feedback and
Update count = count + 1 for each feedback;
Step 2: For 1 <= i <=count for each vehicle
If 0 < Vi <11 then categorize V_i as type A
If 10< Vi < 31 then categorize V_i as type B
If 30< Vi < 46 then categorize V_i as type C
If 45< Vi < 51 then categorize Vi as type D
If 50< Vi < 101 then categorize V_i as type E
Step 3: Set counter: count1: = 1;
Set L_1= 1;
For 2 <= i <= count for each Vehicle
For 1 <= j <= count1
Compare the {type(i) , type(j)} present in the lane
If different update count1 = count1 + 1 and
L_i= count1;
Else
L_i = j;
End of loop;
End of loop;
Step 4: Set counter: count2 = count1;
For 1 <= i <= count -1 for each Vehicle
` For 2 <= j <=count for each Vehicle
If type(V_i)= type(j) and Vi < Vj and arrivaltime($V_{i)}$ <= arrivaltime(V_j)
Set t =arrivaltime (V_j) – arrivaltime (Vi);
Set t1 = 0;
Begin loop
Set t1 = t1 + 1;
Set d = Vi x (t + t1);
Set d1 = Vj x t1;
If d1 <= d Set count2 = count2 + 1;
If L_j = 1 then transition will be to 2 - lane;
If L_j = count1 then transition is count1 - lane;
Else
Transition is either L_j - 1 or Lj + 1;
End loop;
End loop;
End loop;

Step 5: Return Number of lanes required = count1;
Number of transitions required = count2;
Step 6: End

Analysis of the Proposed Algorithm (Algorithm I)

- The above algorithm is implemented on an open unplanned Area.
- The objective will follow linear queue as long as speed/value/cost of proceeding is greater than the immediate next.
- Transition/Cross over are used and they again follow appropriate data structure in order to maintain the preceding step rule.
- Here we assume the lanes are narrow enough to limit the bi-directional approach.
- Here we maintain optimize speed for each lane.
- Here we also maintain the transition points if speed/value/cost of a vehicle is found unable to maintain the normal movement and transition in all the calculated lanes.
- Transition points are recorded with their position and number and it follows appropriate data structure in order to maintain the record.

3.3.2 Algorithm II

Description of the Proposed Algorithm. The primary sections of the proposed algorithm and their major functionalities are described below.

- Step 1. Take relevant information from sensors, i.e. the current speed of the vehicle, arrival time etc. and count the number of vehicles the sensor has entered along with that consider number of lanes that are present in the traffic.
- Step 2. Assign lanes to different vehicles having different current speeds at any time instant t in order to categorize them.
- Step 3. Determine whether the current speed of the vehicle is equal to the speeds present in speed buffers of lanes or not.
- Step 4. This step finds the lane, where, the difference between the vehicle's current speed and lane's speed buffer's average speed is minimum and takes the vehicle to the lane, categorizes it same as the lane's other vehicles, increases the population of the lane, and stores the vehicle's current speed in the speed buffer of the lane.
- Step 5. This step is used for checking total numbers of transitions, i.e. at which point of the lane and from which lane to where the transition will occur, thereby calculating the average speed of the lanes.

Pseudo Code (Algorithm II)

INPUT: Vehicle's name, current speed, arrival time.
OUTPUT: Vehicle's Type, Number of transitions.

Step 1.1: Set count=1; /*used to count the number of vehicles*/
Step 1.2: get_input ()/*Enter the inputs when speed of the vehicle is non-zero. */
Step 1.3: Continue Step 1.1 until sensor stops to give feedback.
Step 2: Set type(1)='A', Enter V_1 into 1st lane's speed buffer, Set 1st lane's population (count_l) as '1', Set n=2.
For 2≤i≤count
Set a buffer buf=0
Loop1 until lane='0'
Loop2 for 1≤j<I for each vehicle
If Vi = Vj
Set buf =1, type(i)=type(j)
Goto Step 3 and send Vi to Step 3 as 'speed1'.
Step 2.1
If buf=1 then end Loop1
If buf=0
Enter Bn=Vi, Set count_l=1, Set type(i)=A++;
End Loop1 /*Bn=n lane speed buffer*/
If lane=0
Then end Loop1.
If lane=0
Then end Loop.
Store buf2=i+1
Step 3: For 1≤i≤lane_l for each lane
If Bi's 1st speed=speed1
Update count_li++;
Set Bi, count_1=speed1
goto step 2.1
Step 4: for buf2≤i≤count
Set c=1, switch=0.
Set min=|Vi, Lc|, /*Lc=c lane's average speed*/
type(i)=1st lane's vehicle type
For 1≤j≤lane_l
Set d=|Vi, Lj|
If d=0
Set type(i)= type(L_j)
Update (j) lane's count_l= (j) lane's count_l+1
Set switch=1
End Loop
If d<min
Then min=d
Set type(i)= type(L_j)
Update (j) th lane's count_l= (j) lane's conut_l+1
Set (j) th lane's speed buffer [count_l] = (i) vehicle's speed (Vi)
If switch=0
Update L1, count_1 ++;
Step 5: Set count2 as count2 =1

For $1 \leq i \leq count-1$
For $2 \leq j \leq count$
If type(i)= type(j) and $Vi < Vj$ and (i) vehicle's arrival time$\leq$ (j) vehicle's arrival time
Set t=(j) vehicle's arrival time - (i) vehicle's arrival time
Set t1=0
Begin loop
Set t1=t1+1
Set d=Vi*(t+t1)
Set d1=Vj*t1
If $d1 \leq d$ set count2 = count2+1
If $L_j = 1$ then transition will be to 2-lane
If L_j =count1 then transition will be to count1-lane
Else transition will be to L_j -1 or L_j +1
End loop
End loop
End loop
For $1 \leq m \leq lane\_1$
Calculate each lane's average speed from its speed buffer.
Step 6: Return Number of transitions required= count2
Step 7: End.

Analysis of Algorithm II: The salient points and features of the proposed algorithm may be analyzed as follows.

- The above algorithm is implemented on an open lane area.
- The objective will follow linear queue as long as speed/value/cost of proceeding to greater than the immediate next.
- Transition/Cross over are used and they again follow appropriate data structure in order to maintain the preceding step rule.
- Here we assume the lanes are narrow enough to limit the bidirectional approach.
- Here we also maintain the transition points if speed/value/cost of a vehicle is found unable to maintain the normal movement and transition in all the calculated lanes.
- Transition points are recorded with their position and number and it follows appropriate data structure in order to maintain the record.

4 Experimental Results and Observations

The optimization of the speed in rush hour traffic with the swarm intelligence approach in an open lane area used the population information as a knowledge base. Primary objective of this approach is to improve the traffic movement in rush hours and to optimize the speed of the vehicles using the concept of transition points between adjacent Lanes.

Proposed algorithms have been implemented using programming language ANSI C in an open platform, on an Intel Pentium IV processor with 1GB physical memory.

Below is the simulated graphical analysis of experimental results thereby obtained

4.1 Simulated Graphical Analysis of the Proposed Algorithms

By implementing the above proposed algorithm and doing the simulation we were able to generate the following graphical results shown in figures 2 and 3 as follows.

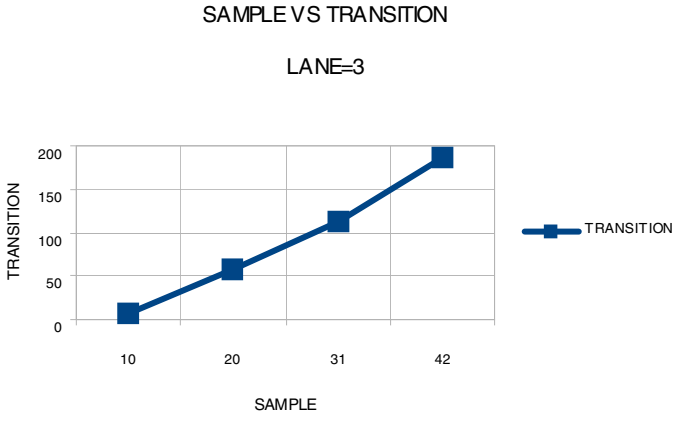


Fig. 2. This figure shows the variation of number of transitions with the number of lanes for a fixed number of samples i.e. 20

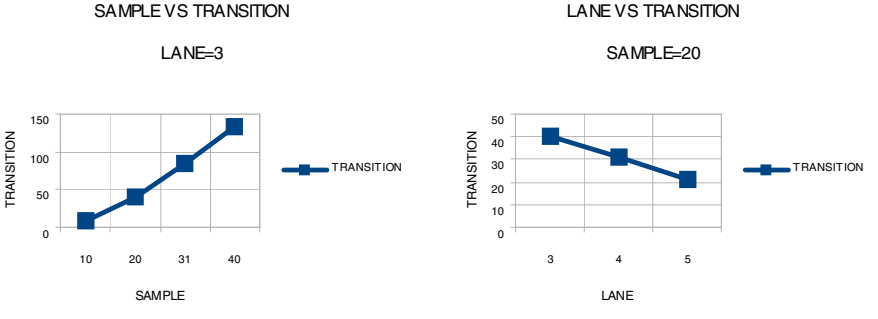


Fig. 3. First figure shows the nature of variation of the number of transitions with the variation of sample size without the population consideration. Second figure shows the same variation with population consideration.

Brief Analysis of Figure 2 and Figure 3

Analysis of the above-simulated results may be interpreted as follows.

- From figure 2 it is clear that as we increase the number of available lanes for a fixed number of samples, the number of transitions is decreasing drastically, which is, very much in conformity with the real life scenario.

- Another important point may be noticed from figure 3. As we are using the population knowledge base, there is a significant improvement in the number of transitions with the result when we were not using population knowledge base under consideration. This shows the effectiveness of our algorithm.

5 Conclusions and Future Scope

The article presented through this paper mainly emphasize on optimal usage of lanes using population information as knowledge base, but at the cost of transitions, because in real life scenario transitions may be too high, hence our future effort will be certainly in this direction.

In this article amount of time taken to transit between lanes has been considered as negligible. However cumulative sum of transition time between lanes in real world problem contributed much in optimality of the proposed solution.

Bio inspired algorithms (like swarm intelligence) has been used with population information as knowledge base, but partial modification of the stated concept taking weighted average of transition information as well as population information will certainly be taken into consideration during implementation and formulation of algorithms in future, there by optimizing various aspects of traffic movement in real world.

References

1. Kononov, J., Bailey, B., Allery, B.K.: The relationship between safety and congestion. Journal of the Transportation Research Board, No. 2083
2. Differentiated speed limits. In: European Transport Conference Differentiated Speed Limits (2007)
3. Messer, C.J., Fambro, D.B.: Critical lane analysis for intersection design. Transportation Research Record (644), 26–35 (1977)
4. Ghosal, P., Chakraborty, A., Das, A., Kim, T.-H., Bhattacharyya, D.: Design of Non-accidental Lane. In: Advances in Computational Intelligence, Man-Machine Systems and Cybernetics, pp. 188–192. WSEAS Press (2010)
5. Ghosal, P., Chakraborty, A., Banerjee, S.: Design of Efficient Knowledge Based Speed Optimization Algorithm in Unplanned Traffic. IUP Journal of Computer Sciences (in press)

Automation of Regression Analysis: Methodology and Approach

Kumar Abhishek[1], Prabhat Kumar[2], and Tushar Sharad[3]

[1] Department of computer Science and Engineering, NIT Patna-800005
kumar.abhishek@nitp.ac.in
[2] Department of Information Technology, NIT Patna-800005
prabhat@nitp.ac.in
[3] Business Technology Analyst, Deloitte India
tushar.sharad@gmail.com

Abstract. Test automation is widely used in different practices all over the world often to save time or to reduce manual effort. However, regression analysis consists of mundane tasks that are performed by software engineers on a daily basis. Automation of a regression analysis raises our hopes by promising a reduction in time and effort, yet at the same time it continues to create as many problems as it has solved. Thus the solution has to take into account the limitations of automation yet reap the maximum benefits. This paper will focus on: The paradigm to be followed while developing automation and the Advantages and Limitations that accompany the process of automating regressions.

Keywords: Regression, Automation.

1 Introduction

A lot of advancement has been made in the field of software development which makes an assortment of tools available for our purposes of coding [1]. What this essentially does, is that the time taken to code a certain program is drastically reduced whereas the testing and product validation teams have a harder job of combing through large amounts of code in shorter periods of time[1-2]. There is a need to improve the testing process by increasing its efficiency and productivity. This can be achieved by using techniques such as automation. This paper details the processes required to identify and diagnose problems faced during automation of regressions and how to tackle them while ensuring reliability and consistency.

Let us take an example: In typical software development/testing company there are at least 5-10 different branches on which concurrent development occurs. Now for each such branch the importance of checking and rechecking the code is immense. So regressions are used which are basically a collection of thousands of test cases each validating and testing some functionality or code in the tool which are run on a

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 481–487.
springerlink.com

daily/weekly basis. The analyses of the results of these regressions are an important part of an engineer's workload since these failures have to be corrected before the next piece of code is checked in so that the tool remains stable. This process of analyzing regressions and assigning failures is tedious and can be automated to reduce man hours spent on it, which would rather be used for other purposes.

2 Need for Automation

Reducing wastage of time [6-7] is one of the top reasons why any test automation is adopted. The moment we try to eliminate the human intervention in a testing process we depend on automation. This places a large question: How reliable are test automations?

Automation also makes the task of error identification and it drastically reduces an engineer's time required to analyze regressions. It is easier to provide detailed reports of all kinds of failures using automation. It would at the same time produce reliable results and increase overall productivity.

As long as there is a need to design a process capable of running in the background in an independent fashion without relying heavily on the programmer and operates reliably, we require automation.

3 Schema for Monitoring Branching Monitoring Process

Without proper planning the development of automation is sure to fail since an evaluation of its pros and cons is a must. So there are a few basic software engineering practices which help us decide whether or not automating a certain process is beneficial. One of the more favored models for development is the waterfall model.

1. Understand the problem statement thoroughly.
2. Determine the language used for scripting (may be shell /Perl) research thoroughly a while before choosing one. The easiest to learn may not be the easiest to implement as you may get to know later [8].
3. Identify your potential areas of doubt (whether those processes can really be automated) before diving into the coding part.
4. Try to modularize the task you perform. The larger task your script is capable of performing, more chances of it being buggy. So we have to be very careful and linking smaller independent scripts may help debug your code later on [10].
5. Introduce certain checks which require human intervention thus ensuring that too much authority and processing power is not bestowed onto the script since they are error prone and have an uncanny tendency to timeout! This could lead to roadblocks in the daily routines of the programmer.

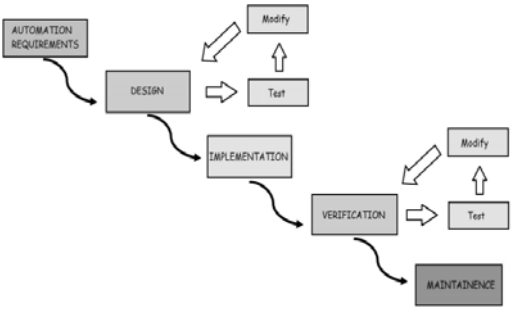


Fig. 1. Model used for developing test automation based on waterfall model

4 Using Automation for Analyzing Regression

The Capabilities of automation should be clearly defined as follows:

1. It should be able to correctly categorize each failure produced during a regression.
2. It should cover all possible reasons and types of failures that occur or may occur during the build.
3. It should have the ability send a comprehensive list of failures that have been caused because of a code checking by the user via email.
4. It should be able to pinpoint the exact reason and handle test cases that need rerun or those that are spurious.
5. No user interaction required. All the data is picked up from last night's regression run. The User is at the receiving end of this automation and he receives an email detailing the failures caused due to his code checking.

A sequence of DFD's will help to understand how this automation is designed.

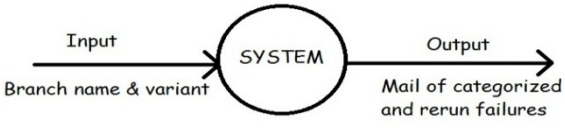


Fig. 2. Overview of the system

There is no guarantee of covering all kinds of errors/failures because while many of them would be dealt with but new kinds of failures introduced by errors in the code would not be detected by the system.

4.1 Constraints

Since the test cases are run on remote machines therefore it makes it virtually impossible to design a script that will continue to run after a remote login attempt has been made as the rlogin shell command transfer control to the new machine.

Logs of each test case run are not present. So the test case would have to be rerun and then examined but this would affect the overall execution time drastically.

Since the job id of each test bench would change every farm run, it is very tough to determine at a later date what might have caused an error. This feature is not implemented in the script. So the reporting of today's failure can occur only today and not sometime in the future! However the benefits in this case outweigh these constraints.

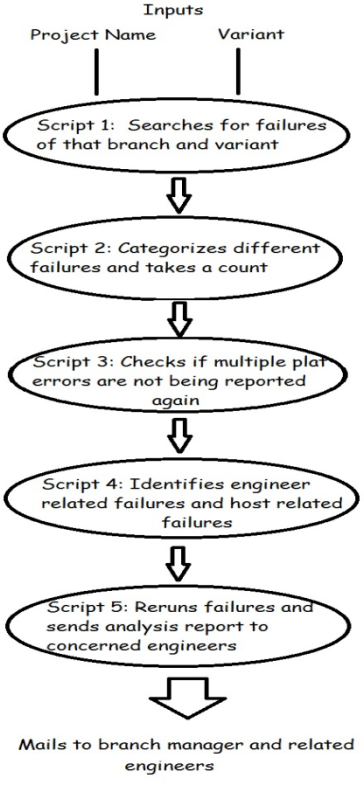


Fig. 3. Detailed explanation of the system (modular approach)

4.2 Assumptions

A certain number of facts need to be assumed to be true. In my case I have, made the following assumptions which are usually true regarding regressions

- Each test case that has caused a failure would have a subsequent log file associated with it.
- All failures have been categorized beforehand and no new types of failures could be introduced in the system.

4.3 Dependencies

- The script only works if the regression has been performed and there are build logs for the same.
- Under any circumstances should the script fail it should automatically try to rerun until it passes. This is quite a problem since if you keep trying to rerun it there would a resource deadlock.
- The script should not be local to only one branch and should be able to run on other releases as well. The design should be well documented and written clearly (egoless programming).

5 Disadvantages of Automation

In most cases automation is not able to eliminate the human element because we cannot schedule large processes solely on the basis of the automation scripts. At the end we require human intervention to ensure reliability of the system[3,14].

Another disadvantage of automation in terms of analyzing regressions is that: It is not able to determine whether the change in code is due to a merge carried out by an intermediate individual or due to a code checking. Since assignment of errors largely depends on the checking, the scripts are not able to distinguish between a merge checking and a code checking. You might argue that if the issue is a merge checking then it should be treated as a merge conflict as any errors before carrying out a merge are the sole responsibility of the branch owner. However this will result in an interminable cycle of blame game and we will be no closer to eliminating the error than we previously were. These certain limitations of automation make it important not only to code correctly but also to assign and identify errors correctly. Another important drawback is that the analysis script cannot be made to identify failures occurring due to incorrect source code, it can identify failures in test cases not the correctness of code/functionality that is tested by it. This is a major drawback of automating regression analysis: If the very issue of identifying code checking errors is not resolved then why the need for automation [7]??

While answering this question we need to remember that an average regression consists of lakhs of test cases. So if there are thousands of errors due to tool failures or missing disk space etc (basically any host/tool based error) we can easily identify them and rectify them before the next build breaks [13].

One last minor problem with automation is that automated mails are seldom paid attention to, so it is quite possible that you might have to pursue the matter later for each failure since cron (scripts which run automatically on a farm machine at a stipulated time) mails are rarely taken into serious consideration. This is a problem with automation that I have experienced and may not happen so in your case.

6 Conclusion

Automation of regression analysis is not a simple process; it requires a lot of planning and research. In this paper we have proposed the use of software engineering process

for small test automations using the modified waterfall model in figure 1. The schema for building such a script has been highlighted in figure 3. The correctness of categorization of failures by the automation is found to be 82% [15].

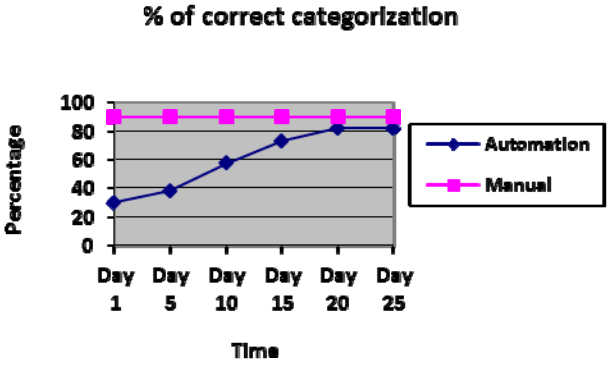


Fig. 4. % of correct categorization (script vs manual)

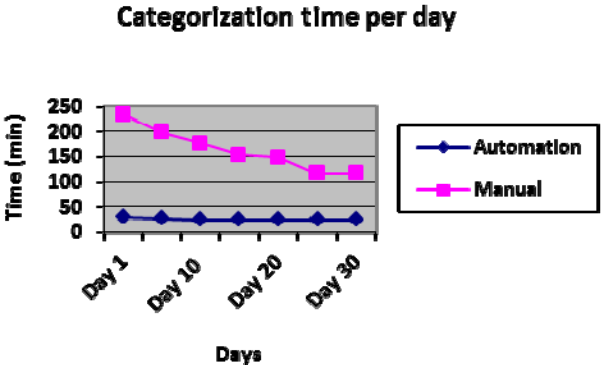


Fig. 5. Time taken to categorize (Automation vs Manual)

These results show the potential use of test automation for analyzing nightly builds for tool testing. "Test automation efforts are prone to being dropped when automators focus on just getting the automation to work. Success requires a more long-term focus [1]. It needs to be maintained and expanded so that it remains functional and relevant as new releases of your product are developed. Concern for the future is an important part of design." So our aim for the future will be to develop a design for automation that minimizes the loop holes and limitations of scripting while at the same time utilizing its expressive and problem solving powers to develop better automation.

References

1. Brett, P.: Seven Steps to Automation Success (June 2001), http://www.io.com/~wazmo/papers/seven_steps.html
2. Bach, J.: Test Automation Snake Oil. Windows Technical Journal (October 1996), http://www.satisfice.com/articles/test_automation_snake_oil.pdf
3. Dustin, E.: Lessons in Test Automation. Software Testing and Quality Engineering (September 1999), http://www.stickyminds.com/sitewide.asp?ObjectId=1802&ObjectType=ART&Function=edetail; Fewster, M., Graham, D.: Software Test Automation. Addison-Wesley (1999)
4. Groder, C.: Building Maintainable GUI Tests. In: Fewster 1999 (1999)
5. Kit, E.: Integrated, Effective Test Design and Automation. Software Development (February 1999), http://www.sdmagazine.com/articles/1999/9902/9902b/9902b.htm
6. Hancock, J.: When to Automate Testing. Testers Network (June 1997), http://www.data-dimensions.com/Testers'Network/jimauto1.htm
7. Hendrickson, E.: Making the Right Choice: The Features you need in a GUI Test Automation Tool. Software Testing and Quality Engineering Magazine (May 1999), http://www.qualitytree.com/feature/mtrc.pdf
8. Hoffman, D.: Heuristic Test Oracles: The Balance between Exhaustive Comparison and No Comparison at All. Software Testing and Quality Engineering Magazine (March 1999)
9. Kaner, C.: Improving the Maintainability of Automated Test Suites. Presented at Quality Week (1997), http://www.kaner.com/lawst1.htm
10. Linz, T., Daigl, M.: How to Automate Testing of Graphical User Interfaces. European Systems and Software Initiative Project No. 24306 (June 1998), http://www.imbus.de/html/GUI/AQUIS-full_paper-1.3.html
11. Jeffries, R.E.: XPractices (1997), http://www.XProgramming.com/Practices/xpractices.htm
12. Marick, B.: When Should a Test Be Automated? Presented at Quality Week (1998), http://www.testing.com/writings/automate.pdf
13. Marick, B.: Classic Testing Mistakes. Presented at STAR (1997), http://www.testing.com/writings/classic/mistakes.html
14. Pettichord, B.: Success with Test Automation. Presented at Quality Week (May 1996), http://www.io.com/~wazmo/succpap.htm
15. Thomson, J.: A Test Automation Journey. In: Fewster 1999 (1999)
16. Weinberg, G.M.: Quality Software Management: Systems Thinking, vol. 1. Dorset House (1992)

A New Hybrid Binary Particle Swarm Optimization Algorithm for Multidimensional Knapsack Problem

Amira Gherboudj, Said Labed, and Salim Chikhi

Computer Science Department, MISC Laboratory,
Mentouri University, Constantine-Algeria
{a.gherboudj,s.labed,Chikhi}@umc.edu.dz

Abstract. In this paper, we presented a New Hybrid Binary Particle Swarm Optimization (NHBPSO). This hybridization consists at combining some principles of Particle Swarm Optimization (PSO) and Crossover Operation of the Genetic Algorithm (GA). The proposed algorithm is used to solving the NP-hard combinatorial optimization problem of Multidimensional Knapsack Problem (MKP). In the aim to access the efficiency and performance of our NHBPSO algorithm we have tested it on some benchmarks from OR-Library and we have compared our results with the obtained results by the standard binary Particle Swarm Optimization with penalty function technique (PSO-P) algorithm and the quantum version (QICSA) of the new metaheuristic Cuckoo Search. The experimental results show a good and promise solution quality obtained by the proposed algorithm which outperforms the PSO-P and QICSA algorithms.

Keywords: Particle Swarm Optimization, Crossover Operation, Multidimensional Knapsack Problem.

1 Introduction

The Multidimensional Knapsack Problem (MKP) is an important issue in the class of knapsack problem. It is a combinatorial optimization problem and it is also a NP-hard problem [6]. In this problem we assume that we have a knapsack with m dimensions, each one has a maximum capacity C_j (j=1,…, m). On the other hand, we have a set n of objects, each object has a profit p_i (i=1,…, n) and a weight w_{ji} in the dimension j of the knapsack. The problem that we want to solve is to find a subset of items that maximize the total profit without exceeding the capacity of all dimensions of the knapsack. The MKP can be formulated as follows:

$$\text{Maximize} \quad \sum_{i=1}^{n} p_i x_i \tag{1}$$

$$\text{Subject to} \quad \sum_{i=1}^{n} w_{ji} x_i \leq C_j \text{ , j=1,…, m} \tag{2}$$

$$x_{id} = \begin{cases} 1 \text{ if the object i is selected} \\ \qquad\qquad i = 1,\ldots\ldots,n \\ 0 \text{ Otherwise} \end{cases} \tag{3}$$

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 489–498.
springerlink.com

The MKP can be used to formulate many industrial problems such as capital budgeting problem, allocating processors and databases in a distributed computer system, cutting stock, project selection and cargo loading problems [5]. Due to its importance and its NP-Hardness, MKP has received the attention of many researches. It was treated by several algorithms for examples, Chu and Beasley [5] proposed a genetic algorithm for the MKP, Alonso et al [3] suggested an evolutionary strategy for MKP based on genetic computation of surrogate multipliers, Li et al [2] suggested a genetic algorithm based on the orthogonal design for MKP, Zhou et al [1] suggested a chaotic neural network combined heuristic strategy for MKP, Angelelli et al [4] proposed Kernel search: A general heuristic for MKP, Kong and Tian [6] proposed a particle swarm optimization to solve the MKP.

In this paper we propose a New Hybrid Binary Particle Swarm Optimization algorithm that we have called NHBPSO, in which we combine some principles of the Particle Swarm Optimization (PSO) and the Crossover operation of the Genetic Algorithm (GA).

The remainder of this paper is organized as follows. The PSO principle is described in section 2. The third section concerns the Binary Particle Swarm Optimization (BPSO) algorithm. In the fourth section we describe the proposed algorithm. Experimental results are provided in section 5 and a conclusion is provided in the sixth section of this paper.

2 PSO Principle

Particle Swarm Optimization (PSO) is a recent metaheuristic. It was created in 1995 by Kennedy and Eberhart [7] for solving optimization problems. It mimics the collective behavior of animals living in groups such as bird flocking and fish schooling. The PSO method involves a set of agents for solving a given problem. This set is called swarm, each swarm is composed of a set of members, they are called particles. Each particle is characterized by position x_{id}= (x_{i1}, x_{i2},…, x_{id},…, x_{iD}) and velocity v_{id}= (v_{i1}, v_{i2},…, v_{id},…, v_{iD}) in a search space of D-dimension. During the search procedure, the particle tends to move towards the best position (solution) found. At each iteration of the search procedure, the particle moves and updates its velocity and its position in the swarm based on experience and the results found by the particle itself, its neighbors and the swarm. It therefore combines three components: its own current velocity, its best position p_{bestid}= (p_{besti1}, p_{besti2},…, p_{bestid},…, p_{bestiD}) and the best position obtained by its informants. Thus the equations for updating the velocity and position of particles are presented below [8]:

$$v_{id}(t)= \omega\, v_{id}\,(t-1) + c_1\, r_1\, (p_{bestid}\,(t-1) - x_{id}\,(t-1)) + c_2\, r_2\, (g_{bestd}\,(t-1) - x_{id}\,(t-1)) \quad (4)$$

$$x_{id}\,(t)= x_{id}\,(t-1) + v_{id}\,(t) \quad (5)$$

ω is an inertia coefficient. (x_{id} (t), x_{id} (t-1)), (v_{id} (t), v_{id} (t-1)): Position and Velocity of particle i in dimension d at times t and t-1, respectively. p_{bestid} (t-1), g_{bestd}(t-1) : the best position obtained by the particle i and the best position obtained by the swarm in dimension d at time t-1, respectively. c_1, c_2: two constants representing the acceleration coefficients. r_1, r_2: random numbers drawn from the interval

[0, 1[. v_{id} (t-1), c_1 r_1 (p_{bestid} (t-1) - x_{id} (t-1)), c_2 r_2 (g_{bestd}(t-1) - x_{id} (t-1)): the three components mentioned above, respectively. A pseudo PSO algorithm is presented and explained in our previous work in [9].

3 BPSO Algorithm

The first version of the Binary Particle Swarm Optimization (BPSO) algorithm (The Standard BPSO algorithm) was proposed in 1997 by Kennedy and Eberhart [10]. In the BPSO algorithm, the position of particle i is represented by a set of bit. The velocity v_{id} of the particle i is calculated from equation (4). v_{id} is a set of real numbers that must be transformed into a set of probabilities, using the sigmoid function as follows:

$$sig(v_{id}) = \frac{1}{1+\exp(-v_{id})} \tag{6}$$

Where sig(v_{id}) represents the probability of bit x_{id} takes the value 1.

To avoid the problem of divergence of the swarm, the velocity v_{id} is generally limited by a maximum value V_{max} and a minimum value $-V_{max}$, i.e. $v_{id} \in [-V_{max}, V_{max}]$. The position x_{id} of particle i is updated as follows:

$$x_{id} = \begin{cases} 1 \text{ if } r<\text{sig}(v_{id}) & r \in [0,1[\\ 0 \text{ Otherwise} \end{cases} \tag{7}$$

Two main parameter problems with BPSO are discussed in [14]. First, the effect of velocity clamping in the continuous PSO is opposite of that in the binary PSO (BPSO). In fact, in the continuous PSO the maximum velocity of the particle encourage the exploration, but it limits the exploration in the binary PSO [14]. The second problem is the difficulties with choosing proper values for inertia weight. In fact, $w < 1$ prevents convergence [14].

4 The Proposed Algorithm (NHBPSO)

PSO is a recent population based metaheuristic that has proved its simplicity of implementation, its effectiveness and its very fast convergence [9]. However, the selection and adaptation of the large number of PSO parameters such as: swarm size, inertia coefficient w, acceleration coefficients c_1 and c_2, play a crucial role for good and efficient operation of PSO. On the other hand, PSO may be easlly trapped into local optima if the global best and local best positions are equal to the position of particle over a number of iterations [11].

In order to benefit from all these advantages and escape all these shortcomings, we are inspired by the PSO principle and the crossover operation of the Genetic algorithm which allows a good exploration of the search space. Our objective is to propose a New Hybrid Binary Particle Swarm Optimization algorithm that provides a

good balance between exploitation and exploration of the search space. The proposed algorithm is described below.

4.1 Representation

The main advantage of the proposed algorithm is that is can be applied to solve all types of optimization problems (continuous, discrete and discrete binary optimization problems) by the appropriate choose of the population type. In fact, since the final solution of the binary particle swarm optimization is a binary solution and the multidimensional knapsack problem is a 0-1 optimization problem, it is an obvious choice to represent and initialize the population with a binary representation. In this aim, we have ulitized binary vectors of size D to represent different particles. The representation of particle i is as follows:

$$x_{id} = [x_{i1}, x_{i2}, \ldots, x_{id}, \ldots, x_{iD}]$$

$$\text{Where} \quad x_{id} = \begin{cases} 1 \text{ If the object is selected} \\ 0 \text{ Otherwise} \end{cases}$$

4.2 Particle Repair (PR) Algorithm

In the MKP, the solution must verify the m constrained of the knapsack to be accepted as a feasible solution. If a solution i exceed the capacity of any dimension of the knapsack, it is considered as infeasible solution and it is not accepted. To repair a solution i, we have proposed Particle Repair algorithm (PR). The PR algorithm allows conversion of an infeasible solution to feasible solution. A pseudo Particle Repair algorithm is presented in Algorithm 1.

Algorithm 1 Particle Repair (PR)

Input : solution vector x
Output : repaired solution vector x

Calculate $R_j = \sum_{i=1}^{n} w_{ji} x_i$, j=1,…, m;

```
For (j=1,…, m){
   While  Rj > Cj{
      Select randomly  i∈ {1 , . . . , n}
      If x_i =1 {
                  x_i = 0;
                  Calculate R_j = Σ_{i=1}^{n} w_ji x_i , j=1,…, m
       }
    }
}
```

Algorithm 2 Crossover Algorithm

Input : Tow particles x_1 and x_2
Output : One particle x_i

1. Choose a step p
2. Choose two random positions c_1 and c_2 from x_1 and x_2 : $c_1, c_2 \in \{1, \ldots, D\text{-}p\}$
3. Swap elements of x_1 from c_1 to c_1+p with those of x_2 from c_2 to c_2+p
4. Swap elements of x_1 from c_2 to c_2+p with those of x_2 from c_1 to c_1+p
5. Calculate the fitness of new particles and select the best one.

4.3 Crossover Operation

Crossover operation is one of the genetic algorithm operations which has introduced by John Holland in 1960 [13]. The main role of the crossover operation is to produce a new population (individual). It consists in combining the characteristics of two individuals (parents) to produce one or two new individuals (children). In the proposed algorithm and in the aim to produce a new population, we have used the crossover operation between the best position p_{bestid} of particle i and its current position x_{id} to produce a new child. This one (i.e. the new child) is crossed with the best position obtained by the swarm g_{bestd} to produce a new particle of the new population. In all cases, we assume that we have 2 particles x_1, x_2 and we want to cross them. We begin by initializing the step p, 2 random positions c_1 and c_2, then we follow steps presented in Algorithm 2 and explained with an example in Fig. 1. Where Algorithm 2 represents a pseudo Crossover Algorithm and Fig. 1 represents an example of Crossover operation. The proposed crossover algorithm gives birth to two new children. To choose which one will represent the new particle, we calculate the fitness f (x_i) of each child and we select the best one.

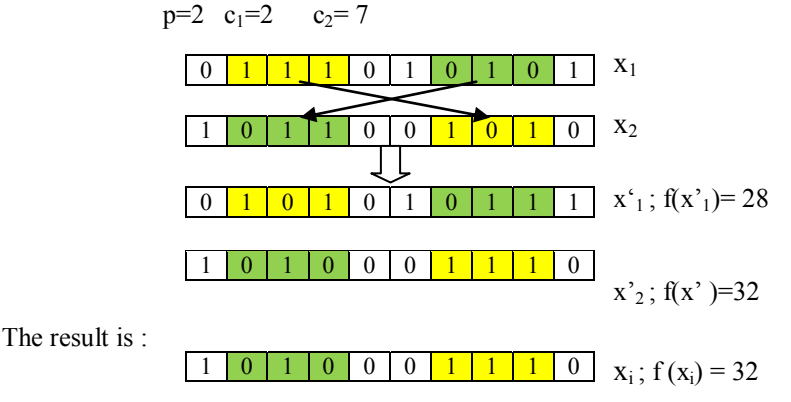


Fig. 1. An example of crossover operation

4.4 Outlines of the Proposed Algorithm

As any algorithm, the first step in the NHBPSO algorithm is to initialize some necessary parameters for good and efficient operation of the algorithm. The main characteristic of the NHBPSO algorithm is its simplicity. In fact, there are fewer parameters to be set in the NHBPSO comparing with other population based metaheuristics such as PSO and GA.

Steps of the proposed algorithm are presented below.

Step1: Initialize a swarm size S and random position of each particle. For each particle, let $p_{bestid} = x_{id}$;
Step2: Apply **PR** algorithm on each infeasible solution and evaluate the fitness of particles;
Step3: Calculate the g_{bestd} ;
Step4: Calculate the new x_{id} of each particle using the following equations:

$$cx_{id} = p_{bestid} \otimes x_{id} \tag{8}$$

$$x_{id} = g_{bestd} \otimes cx_{id} \tag{9}$$

Where the «⊗» operator is the crossover operation of the Genetic algorithm. A pseudo code of the proposed crossover operation is presented in **Algorithm 2**;
Step5: Apply **PR** algorithm on each infeasible solution and evaluate the fitness of particles;
Step6: update p_{bestid} and g_{bestd} as follows:

If $(f(x_{id}) > f(p_{bestid}))$ $p_{bestid} = x_{id}$;
If $(f(p_{bestid}) > f(g_{bestd}))$ $g_{bestd} = p_{bestid}$;

Step7: Stop iterations if stop condition is verified. Return to Step4, otherwise.

The solution of the problem is the last g_{bestd}.

5 Experimental Results

The proposed NHBPSO algorithm was implemented in Matlab 7. To assess the efficiency and performance of our algorithm, we have tested it on some instances from OR-Library [12]. Two parts of experiments were performed. First, we have tested the NHBPSO algorithm on some small size MKP instances taken from different benchmarks named mknap1, HP, SENTO, WEING and WEISH. Moreover, we have tested the NHBPSO algorithm on some big size MKP instances taken from benchmarks named mknapcb1 and mknapcb 4. We have used 5 tests of the benchmarks mknapcb1 (5.100) which have 5 constraints and 100 items, and we have used 5 tests of the benchmarks mknapcb 4 (10.100) which have 10 constraints and 100 items. The obtained results are compared with the exact solution (best known), the obtained solution by PSO-P [6] and QICSA [15] algorithms. Where PSO-P is the standard binary PSO with penalty function technique and QICSA is the quantum version of the new metaheuristic Cuckoo Search [16].

Table1 shows the experimental results of our NHBPSO algorithm with some easy instances taken from the literature. The first column, indicates the instance name, the second and third columns indicate the problem size i.e. number of objects and number of knapsack dimensions respectively. The fourth column indicates the best known solution from OR-Library. Culumn 5 indicates the best results obtained by the NHBPSO. Table1 shows that the proposed algorithm is able to find the best known result of all instances.

Table 2 and Table 3 show the experimental results with some hard instances of mknapcb1 and mknapcb 4. Table 2 shows a comparison in terms of best and average between our NHBPSO algorithm and PSO-P algorithm. The first column indicates the benchmark name. The second column indicates the problem size, i.e. number of objects and number of knapsack dimensions. Column 3 and 4 record the best and average (Avg) results obtained by the NHBPSO and PSO-P during 30 independent runs for each instance. In terms of best and average, Table 2 shows that the proposed algorithm gives better results compared with the PSO-P algorithm which is based on a penalty function technique to deal with the constrained problems.

Table 1. Experimental results with some small size instances

Instance	n	m	best known	NHBPSO
mknap11	6	10	3800	3800
mknap12	10	10	8706,1	8706,1
mknap13	15	10	4015	4015
mknap14	20	10	6120	6120
mknap15	28	10	12400	12400
mknap16	39	5	10618	10618
mknap17	50	5	16537	16537
HP1	28	4	3418	3418
HP2	35	4	3186	3186
PB5	20	10	2139	2139
PB6	40	30	776	776
PB7	37	30	1035	1035
SENTO1	60	30	7772	7772
SENTO2	60	30	8722	8722
WEING1	28	2	141278	141278
WEING2	28	2	130883	130883
WEING3	28	2	95677	95677
WEING4	28	2	119337	119337
WEING7	105	2	1095445	1095445
WEISH01	30	5	4554	4554
WEISH06	40	5	5557	5557
WEISH10	50	5	6339	6339
WEISH15	60	5	7486	7486
WEISH18	70	5	9580	9580
WEISH22	80	5	8947	8947

Table 2. Experimental Results with mknapcb1 and mknapcb 4 instances obtained by NHBPSO and PSO-P

Benchmark Name	Problem Size	NHBPSO		PSO-P	
		Best	Avg	Best	Avg
mknapcb1	5.100.00	23936	23549	22525	22013
	5.100.01	23827	23475	22244	21719
	5.100.02	23234	22921	21822	21050
	5.100.03	23032	22722	22057	21413
	5.100.04	23652	23169	22167	21677
mknapcb 4	10.100.00	22687	22260	20895	20458
	10.100.01	22256	21804	20663	20089
	10.100.02	21744	21233	20058	19582
	10.100.03	22341	21920	20908	20446
	10.100.04	22204	21844	20488	20025

Table 3 shows a comparison in terms of best solution between the exact solution (best known), our NHBPSO algorithm and the QICSA algorithm. The first column indicates the benchmark name. The second column indicates the problem size. The third and fourth Columns record the best results obtained by the NHBPSO and the QICSA algorithms. In terms of best solution, Table 3 shows that the obtained results by the proposed algorithm are nearest to the exact solution compared with those obtained by QICSA algorithm.

Table 3. Experimental Results with mknapcb1 and mknapcb 4 instances obtained by NHBPSO and QICSA

Benchmark Name	Problem Size	Best known	NHBPSO	QICSA
mknapcb1	5.100.00	24381	23936	23416
	5.100.01	24274	23827	22880
	5.100.02	23551	23234	22525
	5.100.03	23534	23032	22727
	5.100.04	23991	23652	22854
mknapcb 4	10.100.00	23064	22687	21796
	10.100.01	22801	22256	21348
	10.100.02	22131	21744	20961
	10.100.03	22772	22341	21377
	10.100.04	22751	22204	21251

Experimental results show that the NHBPSO algorithm gives good and promising results compared with the found results by the PSO-P and QICSA algorithms. These results are very encouraging. They prove the efficiency of the proposed algorithm.

6 Conclusion

In this paper, we have proposed a new hybrid binary particle swarm optimization algorithm that we have called NHBPSO. In the NHBPSO, we have combined some principle of the particle swarm optimization and the crossover operation of the Genetic algorithm. In the aim to verify and prove the efficiency and the performance of our new algorithm, we have tested it on some MKP benchmarks taken from OR-Library. In the first part of experiments, we have compared our results with the best known solution on some small size instances. Moreover, we have compared our results with the obtained result by the PSO-P and QICSA algorithms on some big size instances. Experimental results show a good and encouraging solution quality obtained by our proposed algorithm. Based on these promising results, our fundamental perspective is to integrate some specified knapsack heuristic operators utilizing problem-specific knowledge in the aim to enhance the performance of the proposed algorithm to solve the MKP.

References

1. Zhou, Y., Kuang, Z., Wang, J.: A Chaotic Neural Network Combined Heuristic Strategy for Multidimensional Knapsack Problem. In: Kang, L., Cai, Z., Yan, X., Liu, Y. (eds.) ISICA 2008. LNCS, vol. 5370, pp. 715–722. Springer, Heidelberg (2008)
2. Li, H., Jiao, Y.-C., Zhang, L., Gu, Z.-W.: Genetic Algorithm Based on the Orthogonal Design for Multidimensional Knapsack Problems. In: Jiao, L., Wang, L., Gao, X.-b., Liu, J., Wu, F. (eds.) ICNC 2006, Part I. LNCS, vol. 4221, pp. 696–705. Springer, Heidelberg (2006)
3. Alonso, C.L., Caro, F., Montaña, J.L.: An Evolutionary Strategy for the Multidimensional 0-1 Knapsack Problem Based on Genetic Computation of Surrogate Multipliers. In: Mira, J., Álvarez, J.R. (eds.) IWINAC 2005, Part II. LNCS, vol. 3562, pp. 63–73. Springer, Heidelberg (2005)
4. Angelelli, E., Mansini, R., Speranza, M.G.: Kernel search: A general heuristic for the multi-dimensional knapsack problem. Computers & Operations Research 37, 2017–2026 (2010)
5. Chu, P.C., Beasley, J.E.: A Genetic Algorithm for the Multidimensional Knapsack Problem. Journal of Heuristics 4, 63–86 (1998)
6. Kong, M., Tian, P.: Apply the Particle Swarm Optimization to the Multidimensional Knapsack Problem. In: Rutkowski, L., Tadeusiewicz, R., Zadeh, L.A., Żurada, J.M. (eds.) ICAISC 2006. LNCS (LNAI), vol. 4029, pp. 1140–1149. Springer, Heidelberg (2006)
7. Kennedy, J., Eberhart, R.C.: Particle Swarm Optimization. In: Proc. IEEE Int. Conf. on Neural Networks, WA, Australia, pp. 1942–1948 (1995)
8. Shi, Y., Eberhart, R.: Parameter Selection in Particle Swarm Optimization. In: Porto, V.W., Waagen, D. (eds.) EP 1998. LNCS, vol. 1447, pp. 591–600. Springer, Heidelberg (1998)

9. Gherboudj, A., Chikhi, S.: BPSO Algorithms for Knapsack Problem. In: Özcan, A., Zizka, J., Nagamalai, D. (eds.) WiMo 2011 and CoNeCo 2011. CCIS, vol. 162, pp. 217–227. Springer, Heidelberg (2011)
10. Kennedy, J., Eberhart, R.C.: A discrete binary version of the particle swarm algorithm. In: Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics, Piscatawary, NJ, pp. 4104–4109 (1997)
11. Olamaei, J., Niknam, T., Gharehpetian, G.: Application of particle swarm optimization for distribution feeder reconfiguration considering distributed generators. Appl. Math. Comput. 201(1-2), 575–586 (2008)
12. Beasley, J.E.: OR-Library, http://www.people.brunel.ac.uk/mastjjb/jeb/orlib/mknapinfo.html
13. Holland, J.H.: Adaptation in natural and artificial system. The University of Michigan Press, Ann Arbor (1975)
14. Khanesar, M.-A., Teshnehlab, M., Shoorehdeli, M.-A.: A Novel Binary Particle Swarm Optimization. In: Proceedings of the 15th Mediterranean Conference on Control & Automation, Athens, Greece, July 27– 29, (2007)
15. Layeb, A.: A novel quantum inspired cuckoo search for knapsack problems. Int. J. Bio-Inspired Computation 3(5) (2011)
16. Yang, X.-S., Deb, S.: Engineering Optimisation by Cuckoo Search. Int. J. Mathematical Modelling and Numerical Optimisation 1(4), 330–343 (2010)

A Cooperative Multi-Agent System for Traffic Congestion Management in VANET

Mohamed EL Amine Ameur and Habiba Drias

LRIA, USTHB, Algiers (Algeria)
m_ameur@usthb.dz,
hdrias@usthb.dz

Abstract. Vehicular ad hoc networks (VANETs) are attracting the interest of a great number of academicians and industrials. One of the most interesting features is the possibility to use a spontaneous and inexpensive wireless ad hoc network between vehicles to exchange useful information such as warning the drivers of an accident or a danger. The very recent researches on Vehicular Ad Hoc networks present novel approaches which combine multi-agent technology with transportation systems. In this paper we focus on how to solve the problem of congestion and traffic management through the application of different agent technologies. Having cars equipped with sensors in a VANET, we propose an approach based on multi mobile agent technology. The empirical results have showed the impact of agent and intelligent communications on the Vehicular Ad Hoc networks in reducing the congestion in VANETs.

Keywords: VANET, Routing Protocols, Intelligent Agents, Mobile Agents, NS2.

1 Introduction

The increasing demand for mobility in the 21st century urges researchers from several fields to design more efficient traffic and transportation systems designs, including control devices, techniques to optimize the existing network, and also information systems. A successful experience has been the cross-fertilization between traffic, transportation, and artificial intelligence that dates at least from the1980s and 1990s, when expert systems were built to help traffic experts control traffic lights. [1]

During the last decades, a new paradigm started gaining momentum, known as "Intelligent Agents". In addition, there has been a tremendous progress in traffic engineering based on agent technology. Therefore, traffic and transportation scenarios are extraordinarily appealing for multi-agent technology. Lately, Intelligent Transport System (ITS)[1] has been proposed to improve vehicle safely and comfort of drivers and passengers using large scale wireless techniques such as Vehicular Ad Hoc Network (VANET) [2].

[1] http://www.its.dot.gov/

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 499–508.
springerlink.com

The paper is organized as follows: Firstly, the related works are briefly reviewed. In section 2 we describe the problem of Traffic congestion and control management. The section 3 introduces a variety of technologies related to our study. In the design section, we introduce the architecture we propose composed of three agents types, and show how the proposed system solve the problem of the traffic congestion management in vehicular networks. In section 5, we present the simulation results for the multi-agent architecture. Finally, we conclude the paper with a summary of the achieved outcomes and give directions for future work.

2 Related Works

The works on VANET have been launched by the academic institutions and industrial research laboratories several years ago. Nevertheless, a little research has been dedicated for addressing transportation systems with the agent technology, which seems to be the most adequate according to the geographical distribution of the problem.

In [3], an approach based on using Hitchhiker Mobile Agents for Environment Monitoring is proposed for cars equipped with sensors in a VANET monitoring environment. In [4], a novel data harvesting algorithm for urban monitoring applications called data-taxis is presented. This proposed algorithm has been designed based on biological inspirations. In [5] an intelligent transport system framework based on multi-agent paradigm called CSCW (Intelligent Computer Supported Cooperative work) has been developed. With the proposed technology, drivers of individual vehicles are able to make quick responses to the road emergency. Meanwhile, drivers of individual vehicles around the emergency area can also make the appropriate decision before they reach the road emergency spot.

The fundamental goal of our work is to conceive solutions to enable more robust Multi-agent Systems to cope with the traffic congestion problem in VANET environment. More precisely, we try to develop new network architecture to model the problems of traffic congestion, control and Management in overloaded network environments. This work will stimulate cross disciplinary research in multi agent systems and dynamic Vehicular Wireless Networks for further investigations.

3 Traffic Congestion

Traffic congestion is quickly becoming a spiny problem especially in developed countries. It progressively continues to get worse as the population continues to increase, resulting in an increase in the number of vehicles on the road. There are two ways to combat traffic congestion [6] either by changing road infrastructure to cater to the demands of the road or by developing traffic management systems. Changing the road infrastructure is not an easy process to do as it is very expensive and sometimes it is not possible in the first place because of the increasing number of road lanes and tunnels or bridges building. Also, changing the road infrastructure may improve traffic in one area but make things worse in another one.

Therefore, the other solution would be to develop traffic management systems and driver aids to regulate traffic. Although this solution seems very appropriate, implementing and performing real-tests are very expensive and hard to handle [6, 7].

4 Useful Advanced Technologies

As mentioned previously, our contribution is conceived around two recent technologies which are the VANET technology and the agent software technology. The concepts of both paradigms are presented in the following subsections.

4.1 The VANET Technology

Vehicular Networks are an envision of the Intelligent Transportation Systems (ITS).They are formed when vehicles on the roads are equipped with short range wireless communication devices [8]. VANET is special class of MANET (Mobile Ad-Hoc Network)[2], is established for facilitating communication among nearby vehicles (vehicle-to-vehicle, V2V) and between vehicles to roadside access points (vehicle-to-roadside, V2R) [9, 10].

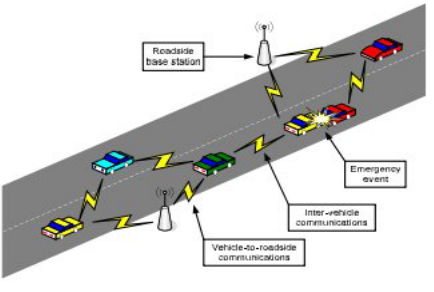


Fig. 1. A VANET consists of vehicles and roadside base stations that exchange primarily safety messages to give the drivers the time to react to life-endangering events [11]

4.2 The Intelligent Agents Technology

Development of agent technology is very much based on research in artificial intelligence and distributed computing. However, attention and practical interest in agents was initiated by the development of network technology. Thus networking is an essential source for developing interest in agents and especially for vehicle networks. At the same time networking itself is an attractive area for agents. The potential of such applications is very high but there are not too many practically successful applications, and the benefits of agents still need to be demonstrated or justified.

[2] http://www.techterms.com/definition/manet

4.3 Mobile Agents

According to its design, a mobile agent can migrate and execute on different machines in a dynamic networked environment. It senses and acts proactively in its environment to realize a set of goals or tasks. Using mobile agents in VANET network, as outlined, provides a number of distinct advantages. Firstly, Mobile agents can migrate from vehicle to vehicle in a network environments .Secondly, they execute asynchronously and autonomously: This is the reason why mobile agents are so promising in wireless networks. Due to the fragile and expensive wireless network connections, a continuous open connection between a mobile device and a fixed network will not be always feasible. In this case the task of the mobile user can be embedded into mobile agents, which can then be dispatched into the fixed network [12]. A third advantage is their dynamic adaptation: mobile agents are capable of sensing their execution environment and take decisions based on that dynamically.

In vehicle network environments, mobile agents can be a good solution as they can be dispatched from a central controller to act locally in the system and thus can respond immediately.

5 The System Design

Usually drivers always choose a shortest way to drive to their destination. However, there will be some problems preventing them from passing through traffic congestion (jams) caused by fatal traffic accidents. Then they need to choose another route.

Our intelligent traffic system based on agent technology provides a solution to the unexpected cases, that is, it enables choosing the best way and taking fast actions to avoid traffic congestion.

In order to simplify the problem, we describe the hypotheses as follows:

- First, make clear the destination and route before departure.
- Each vehicle has a local agent (VEHICL_ AGENT), and contains a mobile agent (MOBILE_ AGENT).
- Each vehicle has the ability to send one or more mobile agents.
- Wireless technology 802.11p. is used in the communication between cars and the base station, as well as among RSU.
- The map is divided to several small-areas, each one contains at least one RSU.
- Each RSU has a local agent (RSU_AGENT), and controls one or more paths.
- A mobile agent may cross several zones to negotiate the right to cross a zone.

5.1 The Roles

If there is a problem of congestion, the cars report a problem by sending mobile agents to the base station near the congestion situation. Before the move of the car to another zone, an agent (Sniffer Agent) is sent by this car to the base station of the new zone from the current RSU, to detect if there is a congestion problem.

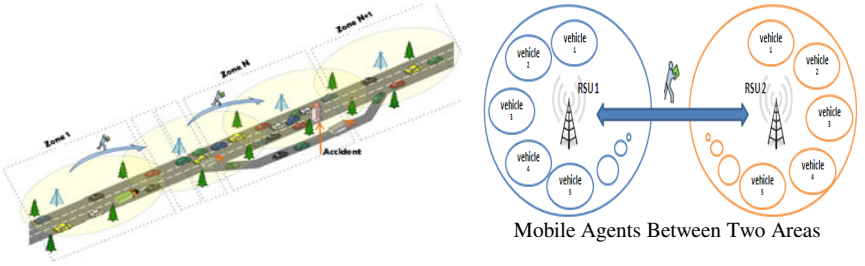


Fig. 2. Migration of Mobile Agent to the Zone Of Relevance ZOR (High way scenarios)

If there is congestion, the car changes its direction before entering the new zone that contains the congestion. If there is no way to avoid congestion (e.g. highway scenarios) the car reduces its speed until it stops in its initial location.

The car starts to test the situation of congestion by sending mobile agents regularly to the new base station until the congestion is unlocked.

5.2 Advantage of Prosing?

At least, we can enumerate the following positive effects:

- The forwarding of the agent in different zone (Sniffer Agent).
- Anticipation of detecting congestion on the emergency region even if no cars exist between the two zones.
- Reduced waiting time of the drivers.
- The agent can be used in the application of comfort such as publicity agents who are sent by hotels, restaurants, fuel stations to cars circulated on the way (or vice versa).

6 Performance Evaluations

6.1 Simulation Setup

We implemented our system using the NS-2 software (version 2.34 with the 802.11p)[3], and in order to generate realistic movement patterns of vehicles we used the SUMO tool[4], that was inputted to NS-2. The TraNS[5] framework uses the TraCI interface to exchange information between SUMO traffic simulator and the NS-2 network simulator and generate realistic simulations of Vehicular Ad hoc NETworks (VANETs). TraNSLite is a stripped-down version of TraNS suitable for quickly generating realistic mobility traces for NS2 from SUMO. This relationship is illustrated in figure 3.

[3] http://www.isi.edu/nsnam/ns/
[4] http://sourceforge.net/apps/mediawiki/sumo/?title=Main_Page
[5] http://trans.epfl.ch/

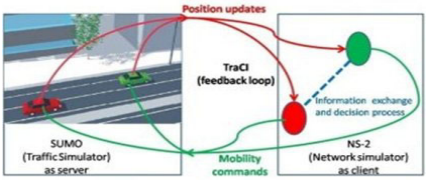


Fig. 3. Real time coupling of Traffic Simulator and Network Simulator using TraCI [9]

In our simulations, Vehicle_agents use TraNS (command-setMaximumSpeed, command-changeRoute) to minimize speed or change direction. Figure 4 shows the map used to generate the movement file and location of RSU, Vehicles move only on the main highway.

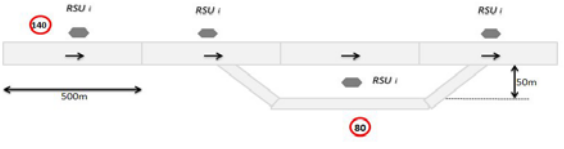


Fig. 4. Map used in the Simulations

The wireless bandwidth and the radio transmission range were assumed 11 Mbps and 100m respectively. Finally, for low-level routing protocol, the *DumbAgent* was employed. The parameters used for simulation are listed in Table 1.

Table 1. Simulations environment

PARAMETER	VALUE
Simulation area	2200 x 200 m 2
Frequency	5.9 GHz
Propagation Models	Two-RayGround model
Maximum Vehicle Speed	140 Km/h
M_agent sending intervals	10 s
Road Warning Messages intervals	5s
congestion duration	200s
mobile agent size	500 octets
Number Of Vehicles	300
Simulation Time	600 s to 1300s

6.2 Performance Evaluation Metrics

Our evaluation considers the following metrics:

Packet Delivery Ratio (PDR): this metric gives the ratio of agents successfully received at the destination and the total number of agents generated, When PDR is

100%, and this means that the destination has received all the agents sent by the source.

Average End-to-End Delay (AE2E): This metric is defined as the duration of time taken by the agent, from its source to its destination (measuring the time taken for the packets arrived at their destination).

Average Travel Time (ATT): This metric is defined as the duration of time taken by vehicle set to reach its destination. The average path duration was a measure of path quality.

6.3 Evaluation Methodology

In our simulations, we assume that accidents happen for the first vehicle that blocks traffic at the third segment, then the vehicle accident and the neighbors vehicles started to send Road Warning Messages to the other vehicles. The duration of congestion is 200s. In order to compare our system with other routing approaches, we simulated three cases (No Congestion No Agent (NCNA), Congestion No Agent (CNA), Congestion and Agent (CA)) to see the impact of use of agents and how agents minimize the Average travel time.

6.4 Simulation Results

We first show some screenshots are taken from simulation. Show that the vehicle agent periodically sends agents to explore the road

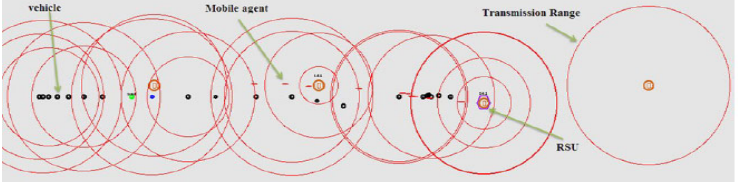


Fig. 5. Mobile agents and travel in the highway

During the simulation, the agents detect congestion (which is located at 1.3km) and warn the local agent of vehicles to change direction and follow the new path

1. **Movement Speed Effect**

In these simulations, we present the results obtained with the change in speed of movement. It varies from 80 to 160, and keeps the other parameters shown in Table 1 unchanged. The speed does not change in the alternate path at 80 km/h.

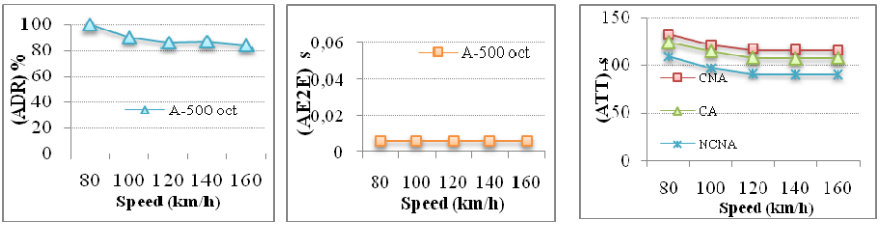


Fig. 6. Metric Performance Under Different Speeds

According to Fig 6-A, we see that ADR decreases progressively as the speed is increases. It is found that no loss agent when the vehicle speed is less than 80 km/s. When we see AE2E, the mobile agent's time propagation in the network remains stable, because our protocol is based on the use of RSU in MA's propagation.

By analyzing the ATT (Fig 6-C), when the average speed increases, the average travel time decreases progressively for the three curves. Because generally when the speed increases the destination travel time reduces. The most important result, as can be clearly seen is that the whole CA curve appears below the NCNA curve. Thus we can conclude that there is always a gain when using agents regardless of the vehicle speed.

2. Vehicles Number Effect

Another important behavior to be tested is one that occurs when multiple vehicles are involved in communication. We vary the number of vehicles in the network from: 100 to 300 vehicles, and keep the other parameters in Table 1 unchanged. We change the size of mobile agent, taking 500 and 1000 octets to see the impact of size in simulation.

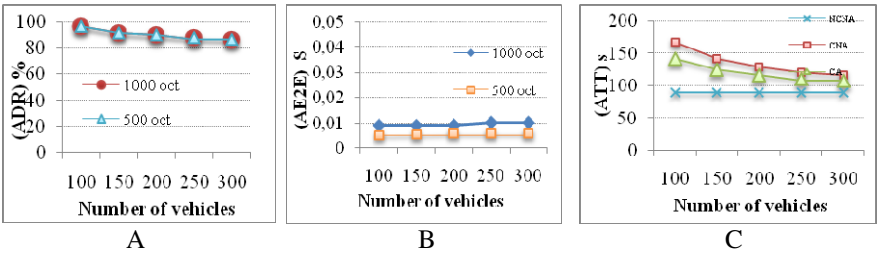


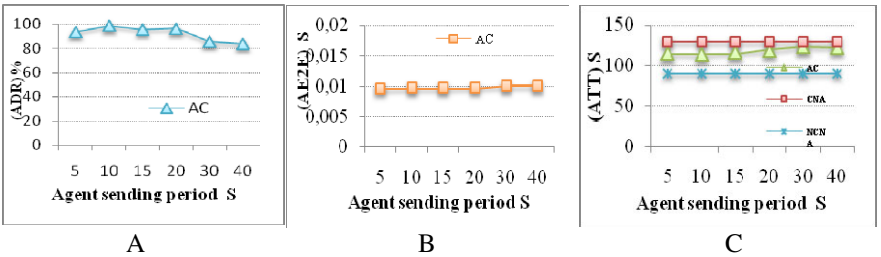
Fig. 7. Metric Performance under Different Number of Vehicles

According to Fig 7-A, we see that the ADR decreases with increasing network traffic. This is explained by the fact that whenever the number of vehicles in the network increases, there is an increasing number of agents lost. The rate of receiving agent then will drop sharply from 99% to 83% for agents of sizes between500 and 1000 octets. Note that the size of mobile agents does not influence the rate of receipt of agent despite the change number of vehicles. With the number of vehicles

between 100 and 150, we can reach a large percentage of agents delivered. By analyzing the E2E, Fig 7-B reveals that the E2E remains almost stable and is not affected by the changes in the number of vehicles. Note that in the same Fig 7-B the A-1000 curve is situated above A-500. This is because when the size of agent augments there is an increase in the number of packets which by the way need more time to move between RSUs. We can say that agents are able to minimize the travel time regardless of the vehicles density.

3. Sending Agent Period Effect

In these simulations, we vary the period of sending agent in the simulation to find the appropriate period that ensures that agents do not load the network. The variations are considered: 5, 10, 15, 20, 30 and 40.



A B C

Fig. 8. Metric Performance Under Different Periods of Sending Agent

From Fig 8-A. we see that the agents' reception rate decreases from 98% to83% when the agent's sending period increases from 20s to 40s. This may be explained by the loss of agents which they take much time to be back to the sending area, whereas the vehicle is already out of it. In Fig 8-B. when we use fixed RSUs, AE2E curve remains stable with the variation of the agent's sending period. In Fig 8-C. we observe the effect of agents use in the simulation expressed by the AC curve versus CNA curve. We conclude that when the sending agent rate is small the vehicle's average travel time is reduced.

7 Conclusion

In this paper, we have designed an intelligent transportation system based on multi-mobile-agent systems and vehicular ad hoc networks. Our approach enables individual vehicle drivers to make quick responses to the problem of congestion and traffic management. We have evaluated the performance of mobile agents in vehicular ad hoc networks using NS-2.

With the proposed technology, drivers of individual vehicles are able to make quick responses to the road congestion. Meanwhile, drivers of individual vehicles around the congestion area can also make the appropriate decision before they reach the road congestion spot.

In the near future, we intend to evaluate the performance of the algorithm in more complex mobility scenarios with communication bottlenecks.

Acknowledgments. The authors express their sincere gratitude to Daniel Krajzewicz, SUMO users, TraNS developers and NS2 users for their collaboration and for invaluable help.

References

1. Bazzan, A., Klügl, F.: Multi-Agent Systems for Traffic and Transportation. In: IGI Global, April 30, pp. xiv (2009)
2. Li, F., Wang, Y.: Routing in vehicular ad hoc networks: a survey. IEEE Vehicular Technology Magazine 6, 12–22 (2007)
3. Urra, O., Ilarri, S., Mena, E., Delot, T.: Using Hitchhiker Mobile Agents for Environment Monitoring. In: Proceedings of PAAMS, March 25-27, pp. 557–566 (2009)
4. Lee, U., Magistretti, E., Gerla, M., Bellavista, P., Lio, P., Lee, K.W.: Bio-inspired Multi-Agent Data Harvesting in a Proactive Urban Monitoring Environment. Elsevier Journal of Ad Hoc Networks 7(4) (June 2009)
5. Fang, Z., Liu, X.: An Agent-Based Intelligent Transport System. In: Shen, W., Yong, J., Yang, Y., Barthès, J.-P.A., Luo, J. (eds.) CSCWD 2007. LNCS, vol. 5236, pp. 304–315. Springer, Heidelberg (2008)
6. Ghazi, A., Ozkul, T.: Design and Simulation of an Artificially Intelligent VANET for Solving Traffic Congestion. MASAUM Journal of Basic and Applied Sciences 1(2), 278–283 (2009)
7. Tumer, K., Agogino, A.K., Welch, Z., Bazzan, A., Kluegl, F.: Traffic Congestion Management as a Learning Agent Coordination Problem. In: Multiagent Architectures for Traffic and Transportation Engineering. Lecture notes in AI, pp. 261–279. Springer (2009)
8. Gainaru, A., Dobre, C., Cristea, V.: A Realistic Mobility Model Based on Social Networks for the Simulation of VANETs. In: Proc. VTC Spring (2009)
9. Bhakthavathsalam, R., Nayak, S., Murthy, G.: Expediency of Penetration Ratio and Evaluation of Mean Throughput for Safety and Commercial Applications in VANETs. In: International Workshop on Communication Technologies for Vehicles, Proceedings of Nets4Cars 2009, Saint-Petersburg Russia, October 13-14 (2009)
10. Baldessari, R., Festag, A., Matos, A., Santos, J., Aguiar, R.: Flexible connectivity management in vehicular communication networs. In: Third Internationl Workshop on Intelligent Transportation, WIT 2006, Hamburg, Germany, pp. 756–758 (2006)
11. Raya, M., Hubaux, J.P.: Securing vehicular ad hoc networks. Journal of Computer Security 15(1), 39–68 (2007)
12. Danny, B., Oshima, L.M.: Seven Good Reasons for Mobile Agents. Communications of ACM 42(3) (March 1999)

An Aspectual Feature Module Based Service Injection Design Pattern for Unstructured Peer-to-Peer Computing Systems

Vishnuvardhan Mannava[1], T. Ramesh[2], and B. Naga Srinivas Repuri[1]

[1] Department of Computer Science and Engineering,
K L University, Vaddeswaram, 522502, A.P., India
vishnu@kluniversity.in, bhaskarr1205@gmail.com
[2] Department of Computer Science and Engineering,
National Institute of Technology, Warangal, 506004, A.P., India
rmesht@nitw.ac.in

Abstract. Adaptability in software is the main fascinating concern for which today's software architects are really interested in providing the autonomic computing. Different programming paradigms have been introduced for enhancing the dynamic behavior of the programs. Few among them are the Aspect oriented programming (AOP) and Feature oriented programming (FOP) with both of them having the ability to modularize the crosscutting concerns, where the former is dependent on aspects, advice and later one on the collaboration design and refinements. In this paper we will propose an Service Injection design pattern for Unstructured Peer-to-Peer networks, which is designed with Aspect-oriented design patterns .We propose this pattern which is an amalgamation of the Worker Object Pattern, Case-Based Reasoning Pattern and Reactor Pattern that can be used to design the Self-Adaptive Systems. Every peer node in the network provides services to many clients, so in order to handle each client request they must be executed in a separate thread which is the functionality provided by the Worker Object Aspect-Oriented Design Pattern. Then after a service request is accepted by a peer server, it will assign the task of serving the request to client with the help of Reactor Pattern. The reactor pattern handles service requests that are delivered concurrently to a peer server by one or more clients. Case-base reasoning pattern is used for composition of services that are provided at different peer servers with the help of JUDDI to serve clients complex service requests. We'll study the amalgamation of the Feature-oriented and Aspect-oriented software development methodology and its usage in developing a design pattern for peer-to-peer networks. So with the help of Aspectual Feature Module technique, we can introduce a new service into the peer system without disturbing the current running code in the server at runtime. In the process of development we also use Java Aspect Components (JAC). A simple UML class diagram is depicted.

Keywords: Autonomic system, Design Patterns, Aspect-Oriented Programming Design Pattern, Feature-Oriented Programming (FOP), Aspect-Oriented Programming (AOP), JXTA, Java Aspect Components (JAC).

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 509–517.
springerlink.com

1 Introduction

As the web continues to grow in terms of content and the number of connected devices, peer-to-peer computing is becoming increasingly prevalent. Some of the popular examples are file sharing, distributed computing, and instant messenger services. Each one of them provides different services, but shares the same mechanism like Discovery of peers, searching, file and data transfer. Currently developed peer-to-peer applications are inefficient with the developers solving the same problems and duplicating the similar infrastructure implementations [1]. Most of the applications are specific to a single platform and can't communicate and share data with different applications.

To overcome the current existing problems Sun Microsystems have introduced JXTA. JXTA is an open set, generalized peer-to-peer (P2P) protocols that allows any networked device –sensors, cell phones, PDA's, laptops, workstations, servers and supercomputers- to communicate and collaborate mutually as peers. The advantage of using the JXTA peer-to-peer programming is that it provides protocols that are programming language independent, multiple implementations, know as bindings, for different environments. The JXTA protocols are all fully interoperable. So with help of JXTA programming technology, we can write and deploy the peer-to-peer services and applications. JXTA protocols standardize the manner in which peers will discover each other, self-organize into peer groups, Advertise and discover network resources, communicate with each other, monitor other.

JXTA overcomes the many of the problems in current existing peer-to-peer systems, some of them are 1) Interoperability – enables the peers provisioning P2P services to locate and communicate with one another independent of network addressing and physical protocols.2) Platform Independent - JXTA provides the developing code with independent form programming languages, network transport protocols, and deployment platforms.3)Ubiquity – JXTA is designed to be accessed by any device not just the PC or a specific deployment platform. In this paper we propose a design pattern for providing the services to peer-clients in unstructured peer-to-peer network.

Design patterns are most often used in developing the software system to implement variable and reusable software with object oriented programming (OOP) [2]. Most of the design patterns in [2] have been successfully applied in OOPs, but at the same time developers have faced some problems like as said in [3] they observed the lack of modularity, composability and reusability in respective object oriented designs [4]. They traced this lack due to the presence of crosscutting concerns. Crosscutting concerns are the design and implementation problems that result in code tangling, scattering, and replication of code when software is decomposed along one dimension [5], e.g., the decomposition into classes and objects in OOP. To overcome this problem some advanced modularization techniques are introduced such as Aspect-oriented programming (AOP) and Feature-oriented programming (FOP). In AOP the crosscutting concerns are handled in separate modules known as aspects, and FOP is used to provide the modularization in terms of feature refinements.

In our proposal of a design pattern for a peer-to-peer system, we use the Aspect-oriented design pattern called Worker Object pattern [6] and Reactor Design Pattern in [7] and Case-Based reasoning [10] Pattern. When comes to the worker object pattern it is an instance of a class that encapsulates a method called a worker method. It will create and handle each client service request in separate thread by making the

job of server easy from looking after each and every client until it completes serving its request .So the server can listen for new client requests if any to handle. The Reactor design pattern is used to handle concurrently more than one client requests for the same service. It does this with the help of a separate event-handler that is responsible for dispatching service-specific requests. Case-Based Reasoning pattern is used for decision-making purpose of deciding the composite service plan to choose for client's complex service request to serve.

2 Related Work

In this section we present some works that deal with unstructured peer-to-peer systems design. There are number of publications representing the design pattern oriented design of the peer-to-peer computing systems. The JXTA protocols standardization provides one of the autonomic computing system properties known as "self-organization" into peer groups. The self-organization is property that provides the autonomic capability in the peer-to-peer design of networks.

V.S.Prasad Vasireddy, Vishnuvardhan Mannava, and T. Ramesh paper [8] discuss applying an Autonomic Design Pattern which is an amalgamation of chain of responsibility and visitor patterns that can be used to analyze or design self-adaptive systems. They harvested this pattern and applied it on unstructured peer to peer networks and Web services environments.

In Sven Apel, Thomas Leich, and Gunter Saake [9] they proposed the symbiosis of FOP and AOP and aspectual feature modules (AFMs), a programming technique that integrates feature modules and aspects. They provide a set of tools that support implementing AFMs on top of Java and C++.

Because of the previous proposed works as described above, we got the inspiration to apply the aspect-oriented design patterns along with inclusion of the feature-oriented software development capability to peer-to-peer computing systems.

3 Proposed Autonomic Design Pattern

One of the objects of this paper is to apply the Aspect-oriented design patterns to the object-oriented code in the current existing application/system. So that a more efficient approach to maintain the system and providing reliable services to the client requests can be achieved. In our proposed Aspectual Feature Module based design pattern for peer-to-peer systems, the client will request for a service to its peer nodes in an unstructured network. This service request is checked initially at the client itself, whether it can be fulfilled by a single service method invocation at a peer server or we have to compose two or more services at different peer servers to serve the clients complex service request. If the case is composition of services, then the Composition Plan is selected by Case-Based reasoning [10] pattern. Composition Plan consists of a new service which is a composition of services at different peer servers that need to be invoked in an ordered fashion to fulfill the current client's request. Then with the help of UDDI Repository the clients get the location details of these composed services at the respective peer servers and then start invoking the services. Then the peer server which check initially, whether the service is currently loaded in the peer's main memory, if not it will load the requested service into the servers main memory with the help of JAC

framework, on the other hand if the service is available with the peer server then it will act as a server, and it will invoke the worker object pattern to handle the client requested service. This pattern also makes sure that if more clients are accessing the same service, then it will permit the request to be handled only until the upper limit on the number of connections that can be served. Once the limit is reached it will ignore the request. So in this way we can reduce the overhead on that peer server by not leading into a blocking or congested state. The request can be handled in another peer who is providing the same service by searching in JUDDI for the service provider details.

Now once the worker object has assigned a separate thread to handle the client request in, then it will call the RequestedServiceHandler method to pass the control to the Reactor design pattern which will look after the service execution. Each service in a peer may consist of several methods and is represented by a separate event handler that is responsible for dispatching service-specific requests. Dispatching of event handlers is performed by an initiation dispatcher, which manages the registered event handlers. Demultiplexing of service requests is performed by a synchronous event demultiplexer. When the service execution is completed the peer-server will return the result to the client.

The important capability that our proposed design pattern provides in the peer-to-peer computing systems that, when a new service is to be added to the peer system in the network without disturbing the running server we can do this with the help of Aspectual Feature Module [9] oriented insertion of the new service into the peer-server code by using the feature refinements property of Feature-Oriented Programming (FOP). So here in order to implement it, we will use the FOP to insert the new service aspectual module as a Feature and Java Aspect Components (JAC) as the mechanism for weaving these new inserted aspectual modules of services in to the current servers environment at run-time. When the any service is no more required in the server's memory, then it can be removed by unweaving the aspectual code with the help of JAC framework [13].

4 Design Pattern Template

To facilitate the organization, understanding, and application of the proposed design patterns, this paper uses a template similar in style to that used in [10].

4.1 Pattern Name

Aspectual Feature Module Based peer-to-peer design pattern

4.2 Classification

Structural-Decision-Making

4.3 Intent

Systematically applies the Aspect-Oriented Design Patterns to an unstructured peer-to-peer Computing System and service injection with a Refinement class for providing new service in the peer-server in terms of a Feature Module.

4.4 Context

Our design pattern may be used when:

- **a)** The applications to which we apply the AOP design patterns will take decisions of which event-handler must be loaded into the server to execute the requested service at run-time.
- **b)** The application will handle each of the client requests in a separate thread by reducing the overhead on the main server thread to get blocked until the first client request is served and making other client requests to get blocked.
- **c)** To include the new service operations into peers as Aspectual Feature Modules [9].

4.5 Proposed Pattern Structure

A UML class diagram for the proposed design Pattern can be found in Fig 1.

4.6 Participants

- **(a) Client:** This application creates JxtaSocket and attempts to connect to JxtaServerSocket. Peer Group will create a default net peer group and a socket is used to connect to JxtaServerSocket. After these steps the client will call the run method to establish a connection to receive and send data. The startJxta method is called in the client to create a configuration form a default configuration and then instantiates the JXTA platform and creates the default net peer group. Once the net peer group is created the client will send a Pipe Advertisement for requesting a service in the peer-to-peer network.
- **(b) Server:** First the default net peer group is created with Peer Group. Creates a JxtaSeverSocket to accept connections, and then executes the run method to accept connections from clients and send receive data. The server will start listening for the service requests, if any of the service requests matches the service list it provides.
- **(c) Connection Handler:** This will take care of the connections with multiple clients and sending and receiving the data between the clients and peer-server.
- **(d) Worker Object Aspect**: Once the connection is established between the server and client, the worker object will create a separate thread for the connected client to run the requested service in a new thread for that it will call the Reactor Pattern method to handle the requested service execution.
- **(e) Initiation Dispatcher:** It defines an interface for registering, removing, and dispatching the event handlers.
- **(f) Synchronous Event Demultiplexer:** The synchronous Event Demultiplexer is responsible for waiting until new events occur. When it detects new events, it will inform the Initiation Dispatcher to call back application-specific event handler.
- **(g) Event Handler:** Specifies an interface consisting of a hook method [11] [7] that abstractly represents the dispatching operation for service-specific events. This method must be implemented by application-specific services.
- **(h) Concrete Event Handler:** Implements the hook method [11] [7], as well as the methods to process these events in an application-specific manner. Applications

register Concrete Event Handlers with the Initiation Dispatcher to process certain types of events. When these events arrive, the Initiation Dispatcher calls back the hook method of the appropriate Concrete Event Handler.

(i) **Handle Event Aspect:** This is the aspect-oriented implementation module that will be viewed into the concrete event handling class, so that only a particular requested service method get viewed into the code at run-time.

(j) **Refines Class Event Handler:** This will add a new event handler for a new service that is inserted into the peer, in such a way that the insertion will be done as a new feature with the help of FOP.

(k) **Trigger_Service:** This class is responsible for accepting the clients service requests and then generating a service trigger to the inference engine to decide whether there is any requirement for the composition of services to handle complex clients service requests.

(l) **Inference Engine:** This class is responsible for the process of finding a perfect service composition plan that can be adapted to fulfill the client's service request. And then it will pass the service request as input to the fixed rules class to make a perfect decision regarding the composition plan selection based on a set of Rules in Rule class.

(m) **Decision:** This returns the client with a perfectly matched Plan to perform the composition of the services to serve the client's complex service requests.

4.7 Consequences

(a) With the use of this pattern we can use the benefits of worker object pattern, where it will handle each service execution in a separate thread by not executing in the Main thread itself.

(b) In a single-threaded application process, Event Handlers are not pre-empted while they are executing. This implies that an Event Handler should not perform blocking I/O on an individual Handle since this will block the entire process and impede the responsiveness for clients connected to other Handles.

(c) We use the Feature-Oriented Programming to insert the new service handler as a feature into the server code.

(d) By the use of dynamic crosscutting concerns of the Aspect-Oriented Programming the system will me executing fast as the decisions are made at run-time.

4.8 Related Design Patterns

(a) Chain of Responsibility [2]: the Reactor [7] associates a specific Event Handler with a particular source of events, whereas the Chain of responsibility pattern searches the chain to locate the first matching Event Handler.

(b) Active Object Pattern [12]: The active object pattern is used when the threads are not available or when the overhead and the complexity of threading is undesirable [7].

4.9 Roles of Our Design Patterns in Unstructured Peer-to-Peer Computing Systems

(a) Worker Object Pattern [6]: The worker object pattern is an instance of a class that encapsulates a worker method. A worker object can be passed

around, stored, and invoked. The worker object pattern offers a new opportunity to deal with otherwise complex problems. It will provide the server with the facility to handle the service request form different clients in a separate per client connection. We may use this pattern in different situations like, implementing thread safety in swing applications and improving the responsiveness of the UI applications to performing authorization and transaction management.

(b) Reactor Pattern [7]: The reactor design pattern handles service requests that are received concurrently from more than one client. In our proposed pattern we use this design pattern for efficiently handling the requested services. Each service in an application may consist of and is represented by a separate event-handler that dispatches the service-specific request. So this task of dispatching is done by initiation dispatcher, which itself manages the registered event handlers.

(c) Case-Based Reasoning Pattern [10]: The case-based reasoning design pattern is responsible for the decision-making process of the selection of a perfect composition plan that should be selected in order to compose the services that are available at different peer servers to serve the clients for their complex service requests.

5 Feature Based Service Insertion into Peer-to-Peer Computing System

The peer-server in the peer-to-peer computing system will need to introduce a new service at particular instance of time into the server. The inclusion of a new service into the peer can be done by inserting the event-handler for that service a new feature with the help of Feature-Oriented Programming. So here we may need to introduce the new service advice into the Aspect that provides the different services, this can be done with the help of Aspectual Feature Module in [9].where the new aspects can be inserted into the system as a feature.

6 Conclusion and Future Work

In this paper we have proposed a pattern to facilitate the ease of developing unstructured peer-to-peer computing systems. So with the help of our proposed design pattern named Service Injection Design pattern for unstructured peer-to-peer systems, provide services to clients with the help of Aspect-Oriented design patterns. So with this pattern we can handle the service-request of the clients and inject the new services into peer-server code as feature modules. Several future directions of work are possible. We are examining how these design patterns can be inserted into a middleware technology , so that we can provide the Autonomic properties inside the middleware technologies like JXTA with the help of Features-Oriented and Aspect-Oriented programming methods.

The view of our proposed design pattern for the unstructured peer-to-peer computing System can be seen in the form of a class diagram see Fig 1.

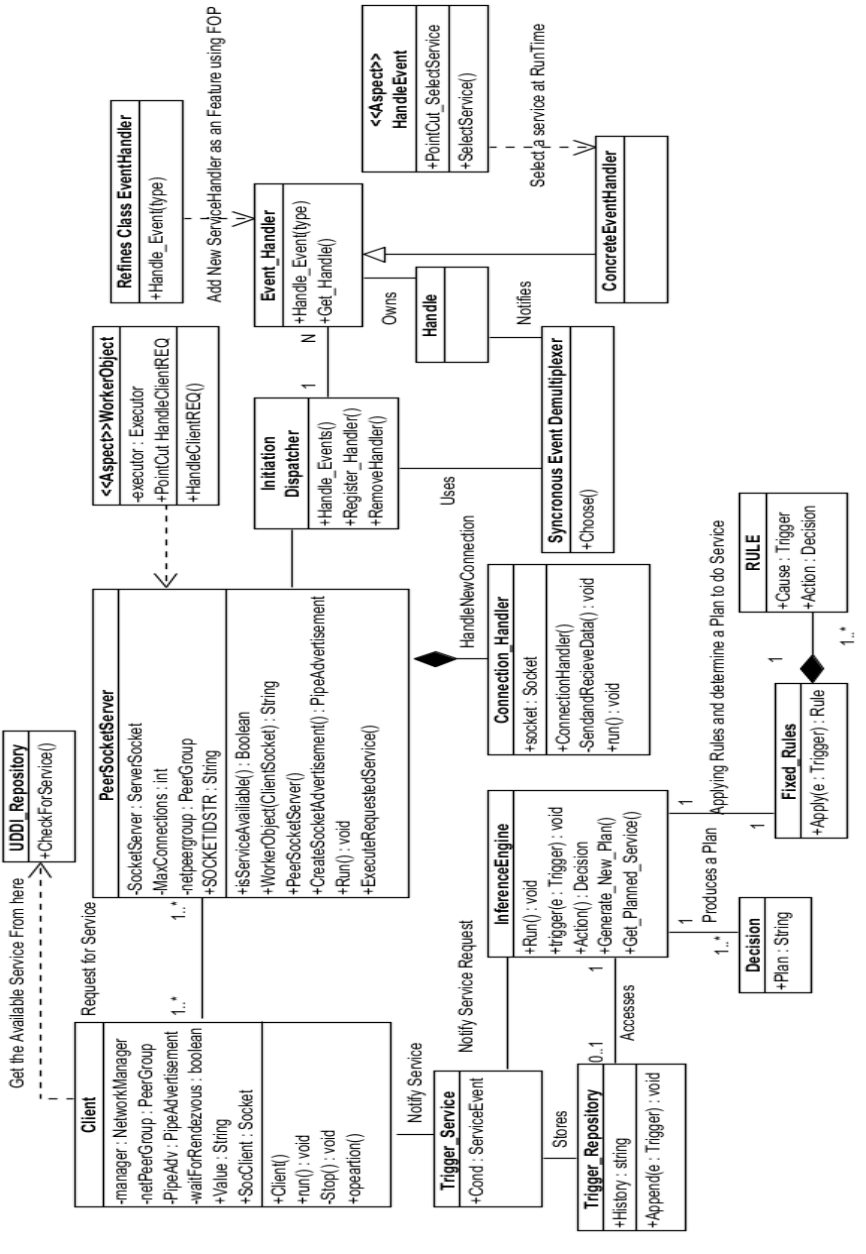


Fig. 1. Applying Design Pattern for the Peer-to-Peer Computing System

References

1. JXTA Java Standard Edition v2.5: Programmers Guide, September 10(2007), 2002-2007 Sun Microsystems, Inc.
2. Gamma, E., Helm, R., Johnson, R., Vlissides, J.: Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley (1995)
3. Kuhlemann, Rosenmuller, M., Apel, S., Leich, T.: On the Duality of Aspect-Oriented and Feature-Oriented Design Patterns. In: Proceedings of the 6th Workshop on Aspects, Components, and Patterns for Infrastructure Software. ACM, New York (2007), doi:10.1145/1233901.1233906
4. Hannemann, J., Kiczales, G.: Design Pattern Implementation in Java and AspectJ. In: Proceedings of the International Conference on Object-Oriented Programming, Systems, Languages, and Applications (OOPSLA), pp. 16–17 (2002), doi:10.1145/583854.582436
5. Tarr, P., Ossher, H., Harrison, W., Sutton, J.S.M.: N Degrees of Separation: Multi-Dimensional Separation of Concerns. In: Proceedings of the International Conference on Software Engineering, ICSE, pp. 107–119 (1999), doi:10.1145/302405.302457
6. Laddad, R.: AspectJ in Action. In: Maning, 2nd edn., ch. 12 (2010) ISBN: 1933988053
7. Schmid, D.C.: Reactor: An Object Behavioral Pattern for Demultiplexing and Dispatching Handles for Synchronous Events. Addison-Wesley (1995)
8. Prasad Vasireddy, V.S., Mannava, V., Ramesh, T.: A Novel Autonomic Design Pattern for Invocation of Services. In: Wyld, D.C., Wozniak, M., Chaki, N., Meghanathan, N., Nagamalai, D. (eds.) CNSA 2011. CCIS, vol. 196, pp. 545–551. Springer, Heidelberg (2011)
9. Apel, S., Leich, T., Saake, G.: Aspectual Feature Modules. IEEE Transactions on Software Engineering 34(2) (2008), doi: 10.1109/TSE, 70770
10. Ramirez, A.J., Betty, H.C.: Cheng Design Patterns for Developing Dynamically Adaptive Systems. In: Proceedings of the 2010 ICSE Workshop on Software Engineering for Adaptive and Self-Managing Systems, ACM, New York (2010), doi:10.1145/1808984.1808990
11. Pree, W.: Design Patterns for Object-Oriented Software Development. Addison-Wesley, Reading (1994)
12. Lavender, R.G., Schmidt, D.C.: Active Object: an Object Behavioral Pattern for Concurrent Programming. In: Proceedings of the 2nd Annual Conference on the Pattern Languages of Programs, Monticello, Illinois, pp. 1–7 (September 1995) ISBN: 0-201-895277
13. Pawlak, R., Seinturier, L., Retaillé, J.-P.: Foundations of AOP for J2EE Development, ch. 8. Apress (2005) ISBN13: 978-1-59059-507-7

A Novel Hybrid Approach to N-Queen Problem

Kavishi Agarwal, Akshita Sinha, and M. Hima Bindu

Jaypee Institute of Information Technology,
A-10, Sector-62
Noida 201 307
Uttar Pradesh, India
{agarwal.kavishi1990,akshita.sinha100}@gmail.com,
hima.bindu@jiit.ac.in

Abstract. These days computers deals with highly intricate problems. This paper also discusses such kind of complex problems called Constraint Satisfaction Problems (CSP). These problems have a set of variables, domain from which a variable takes its value and a set of constraints applied on these variables. For example N-Queen problem, timetabling problem, scheduling problem etc. are CSPs. In practical scenario, it is unlikely to obtain a solution that satisfies all constraints or most of the constraints. Such a solution is an exact solution. Even if an exact algorithm can be developed its time or space complexity may turnout unacceptable. In reality, it is often sufficient to find an approximate or partial solution to such NP problems using heuristic algorithms. Heuristic methods are used to speed up the process of finding a satisfactory solution, where an exhaustive search is impractical; hence resulting in guaranteed and approximate solutions. This paper proposes an efficient hybrid solution for standard N-Queen problem using Ant Colony Optimization and Genetic Algorithm. It also compares the performances of classical backtrack and brute force methods and heuristic methods, Simulated annealing and Genetic algorithm on N-Queen problem.

Keywords: Constraint Satisfaction Problems, N-Queen, Hybridized Heuristic, Ant Colony Optimization (ACO), Genetic Algorithm (GA).

1 Introduction

Problems with no deterministic solutions that run in a polynomial time are called NP Problems. Constraint satisfaction problems also belong to this class.

Constraint satisfaction problems (CSP) s are mathematical problems that can be expressed in the following form. Given a set of variables $\{x_1, x_2 \ldots x_n\}$, for each variable x_i, a domain D_i is available with the possible values for that variable and a set of constraints i.e. relations, that are assumed to hold between the values of variables.

The Classic N-Queen problem is to place 8 queens (N=8), on the 8X8 chessboard such that no two queens attack. This problem can be generalized as placing N non attacking queens on the NXN board. For this problem, the set of variables is the set of N-Queens, the domain is the set of N^2 board positions and the constraint is that no two queens must share the same row or the same column or the same diagonal.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 519–527.
springerlink.com

Heuristic refers to an experience-based technique for problem solving, learning, and discovery. Examples of this method include using a "rule of thumb", an educated guess, an intuitive judgment, or common sense. Heuristic methods are applied on NP problems because in such problems, finding an exact solution is not possible and also not desirable. A method that speeds up the process of finding the satisfactory solution is acceptable in such scenarios. Heuristic Algorithms gives approximate solution. This paper therefore discusses the implementation of Heuristic Algorithms for solving Constraint Satisfaction Problems.

Classical Algorithms can solve the N-Queen problem efficiently only till N<=20. For higher values of N, heuristics are employed to get an efficient solution. The idea behind the design of hybrid algorithms is very simple. It solves the problem by the combination of two different algorithms. Hybrid algorithms exploit the good properties of different methods by applying them to problems they can efficiently solve. A recent paper [1] talks of ACO being used on N-Queen for the first time. This paper proposes a novel hybrid approach for N-Queen problem using *Ant Colony Optimization* and *Genetic Algorithm*. The novelty lies in the hybridization of ACO, in this paper, with GA.

2 Related Work

Previously, a reasonable amount of work has been done on this problem. Classical approaches like Backtracking and Brute-Forcing have been applied in the past. Brute Force is not efficient for the N-Queen problem and has a worst time complexity of the order of O (n!).

Backtracking is a general algorithm for finding some or all solutions to computational problems by incrementally building candidates to the solutions, and abandoning each partial solution as soon as it determines that partial solution cannot possibly be completed to a final valid solution. Complexity of backtracking typically rises exponentially with problem size.

Salabat Khan in [1] applied Ant Colony Optimization for the very first time on the N-Queen problem. The solution they proposed was working efficiently for 8-queen problem and was believed to be capable of easily extended to large values of '*n*' because of the simplistic model of the search space. ACO was concluded to provide better solution in reasonable amount of time for combinatorial optimization problems.

K.D. Crawford in [2, 9] applied Genetic Algorithm and has discussed two ways to solve N-Queen. The experiments for the N-Queen problem discussed in this paper were conducted on various populations and board sizes. The paper investigated the effectiveness of genetic algorithm in searching for good quality solutions.

Also Marko BoiikoviC in [3] discusses a Global Parallel Genetic Algorithm that involves the concept of parallelization. Converting the solution approach into parallel processes can significantly improve their performance. The slaves were enabled to run simultaneous selections and crossovers freeing master process from most tasks (population initialization and mutations during the run were still performed by the master thread).But GPGA is not suitable for massive parallel processing, and shows increase in performance for only a small number of parallel-processing units.

Ivica et al. provided a comparison of different techniques in [4]. These techniques include Tabu Search and GA.

Nguyen Duc Thanh in [5] proposed a hybrid algorithm that combined genetic and heuristic approach. By using this method, solving timetabling problem was converted to finding the optimal arrangement of elements on a 2D matrix.Hence it helped us to conclude that some hybrid approach can be designed for solving N-Queen problem also.

The understanding of the constraint satisfaction problems with respect to these heuristic approaches was developed as discussed in [6, 10].

However we have decided to implement a hybridized technique which combines the evolutionary GA with heuristics like Simulated Annealing to give a better and efficient solution.

3 Present Work

3.1 Problem Statement

This paper aims to provide an efficient heuristic solution to the N-Queen Problem. This paper proposes a novel hybrid algorithm using Ant colony Optimization and Genetic Algorithm. This paper also shows the implementation of Genetic algorithm, Simulated Annealing, Backtracking and brute force algorithms on the N-Queen problem and compares their performances to provide the best solution for the N-Queen Problem.

3.2 Methodology of Solution

Below we have described the algorithms that we have applied on the N-Queen problem and also the hybrid solution proposed by us.

Genetic Algorithm

Genetic algorithm is an evolutionary algorithm that mimics the process of natural evolution. It is based on three biological principles: selection, crossover and mutation. Firstly, the random generation of potential solutions is generated called chromosomes. The chromosomes would be in the form of n tuple (q1, q2…qn) where the Queen Number in the tuple represents the Queen's row position and its position in the tuple represents the Queen's column position as shown in Fig 1. Hence the row and column conflicts are automatically resolved. Only the diagonal conflicts are to be resolved.

Fitness of each individual in the generation is evaluated using the fitness function. Only certain numbers of individuals are selected for the next generation using the selection operator. Selected individuals act as parents. Parents undergo the process of crossover and mutation to produce new generation.

Fitness function used here is the measure of the number of diagonal conflicts for the queen. Greater the number of diagonal conflicts, lesser is the fitness and hence less chances of getting selected into the new generation. Thus, each generation would be fitter than the previous generation.

Selection method employed here is an elite selection method. Crossover is done through Partially Matched Crossover (PMX) method. In PMX, any two elements of the tuple are exchanged as shown in figure 1 and then to prevent forming invalid

tuples, duplicates are removed as shown in Figure 2. Then mutation of the children takes place by randomly swapping the position of any one element in a tuple with the other tuple. But its probability is kept very low. Now this process repeats until the best generation is found. The complexity of the N-Queen reduces to O (n).

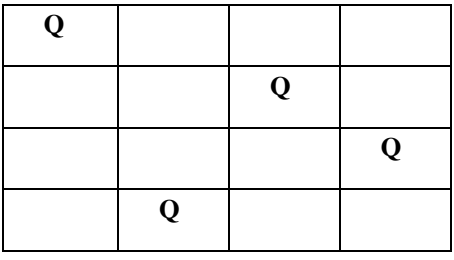


Fig. 1. n-tuple notation example

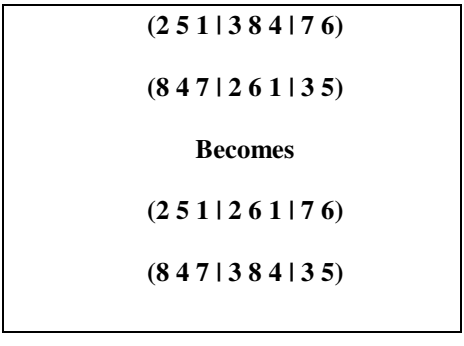


Fig. 2. PMX Crossover step 1

In most cases it will result in invalid tuples since the numbers in the tuple must be unique. Second step in the PMX crossover eliminates duplicates.

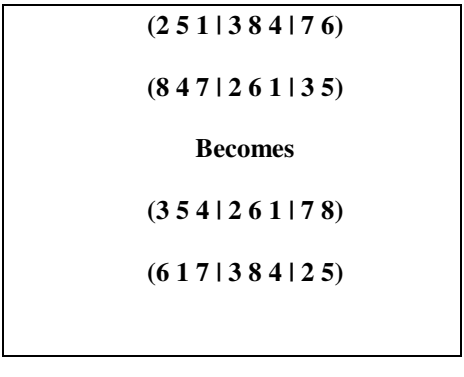


Fig. 3. PMX Crossover step 2

Simulated Annealing

Simulated Annealing (SA) is a generic probabilistic meta-heuristic algorithm for locating a good approximation to the global optimum of a given function in a large search space.

Basically, annealing is a process performed in order to relax the system to a state with minimum free energy. A control parameter is introduced here used to control the transition of the state. The control parameter used is the "temperature". The "temperature" of the system/function to be optimized should have produce a similar effect as the temperature of the physical system. It must condition the number of accessible states and lead to the optimal state, if the temperature is lowered gradually in a slow and controlled manner and should proceed towards a local minimum if the temperature is lowered abruptly.

We check for change in energy (ΔS) here:

if $\Delta S \leq 0$ - modification will be accepted
if $\Delta S > 0$- modification accepted with probability
exp ($-\Delta S/T$)

$\Delta S \leq 0$ denotes that the new state is better than the current state. The cooling ratio to anneal the temperature is denoted by **α.** The value of α varies between 0 and 1. Below we give a pseudo code for the above proposed SA solution:

```
Simulated Annealing(boardsize)
  while( iter < maxiteration and bestCollision !=0)
  check if newstate collision < current state collision
  best state= new state;
  else
  Anneal().
```

Ant Colony Optimization

Ant colony optimization (ACO) is a meta-heuristic that is inspired by intelligent behaviors of ants. We have these ants acting as agents in this multi-agent system. The ants in actual are believed to modify the natural environment by depositing a chemical substance called pheromone. The higher the concentration of pheromone, the better is the path. This path will then guide the ants to find the best solution.

Consider G= (V, E) to be a connected graph, with V representing the vertices of the graph and E represents the edges connecting the vertices. The path length is given by the summation of cost values on edges constituting the path. The pheromone value corresponding to each edge e(i,j) connecting the nodes V_i and V_j will be modified by the ants on visiting these nodes. The decision of movement of ant will depend on the probability associated with that path as in (1).

$$P_{ij} = \frac{[\tau_{ij}]^{\alpha}.[\eta_{i,j}]^{\beta}}{\sum_{k \in S} ([\tau_{ik}]^{\alpha}.[\eta_{ik}]^{\beta})} \quad (1)$$

Here τ_{ij} is the pheromone value on edge(i, j), $\eta_{i,j}$ is a heuristic value calculated as $1/d_{i,j}$. The factors α and β are influencing factors of pheromone value and heuristic value respectively. Values of α and β lies between 0-2.

The value of $d_{i,j}$ when applied to our N-Queen problem will depend on the cost factor associated with the tuples representing a particular solution. The cost value will come out to be the conflicts in that solution.

Some best ants (having good solutions) or all ants modify the pheromone values on the edges added to their tour. One possible modification may be done as in (2):

$$\tau_{i,j} = \tau_{i,j} + Q/\mathrm{N} \tag{2}$$

where Q is any constant and N is the least number of conflicts till now. With time, concentration of pheromone decreases due to diffusion affects; a natural phenomenon known as evaporation.

This also ensures that old pheromone should not have a too strong influence on the future.

This can be done as in (3):

$$\tau_{i,j} = \tau_{i,j} .\rho \ \{\text{where } \rho \text{ will be between 0 and 1}\} \tag{3}$$

where ρ is the evaporation parameter.

ACO and GA Hybrid Approach

This section proposes a hybrid approach using Ant Colony Optimization and Genetic Algorithm for the N-Queen problem. We propose a novel hybrid approach of solving N-Queen problem using Ant Colony Optimization and Genetic Algorithm. According to the diagram in our method, as shown in Fig4.ACO and GA algorithms were implemented consecutively. The ACO system is initialized with the initial parameters as discussed in (3.2.3). The objective functions of the two algorithms are the same as discussed before.

Figure4 shows that ACO works with multiple ants or agents. After its few iterations i.e. when convergence occurs, the initialization process of GA takes the output solutions of ACO as the initial population of chromosomes. It then performs its iterations until the best solution (solution with no conflicts) is obtained.

The hybrid of ACO and GA has never been applied on the N-Queen problem. Moreover, GA is well known to give optimal results for the CSPs but in order to deal with convergence issues of GA, ACO is being used firstly to ensure convergence. The pseudo code for the above approach is discussed below:

1) Generate multiple ant agents;
2) Initialize the pheromone matrix with random initial values;
3) Initialize other ACO parameters;
4) Pheromone local and global update for each iteration;

5) After some iterations, select the solutions generated by the multi-agent system.
6) The refined solutions are taken as the initial population for the GA system.

7) The selection, crossover and mutation operations are s applied as per the GA explained in the section above (3.2.1).
8) Go to step 6 and continue the iterations until the best solution is obtained.

The algorithm proposed above in the pseudo code starts with the generation of multiple ant agents. The pheromone value is then allotted to each agent and thus the corresponding pheromone matrix is generated with initial values. The ACO parameters, α, β, and Q are to be initialized as in (3.2.3). Then after obtaining the solution of ACO, GA starts its execution.

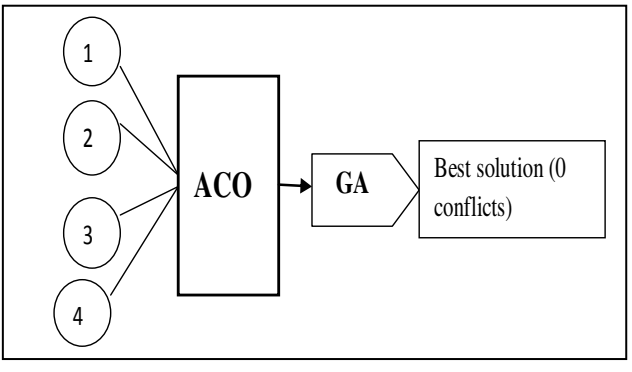


Fig. 4. Hybrid Illustration

4 Applications of N-Queen Problem

Since each solution to the n-queens problem forms a non conflict pattern, the N-Queens problem has many practical,

Scientific and Engineering applications. Some applications are given below.

To achieve high communication bandwidth in a narrowband Directional communication system, an array of *n* transmitters/receivers must be placed without any interference with each other. With n transmitters/receivers placed in a non conflict pattern, which corresponds to a solution to the n-queens problem, each transmitter/receiver can communicate with the outside world freely in eight directions (i.e., two horizontal directions, two vertical directions, and four diagonal directions) without being obscured by other transmitters/receivers.

Further, in preventing deadlock, traffic control, VLSI technology etc N-Queen finds its application.

Table 1. Results for genetic algorithm (N=8)

Population	Generations	crossover Probability	mutation Probability	Execution Time(sec)
100	10	0.7	0.2	.005
200	10	0.7	0.2	.005
500	10	0.7	0.2	.009
1000	10	0.7	0.2	.011
1000	40	0.7	0.2	.014
1000	80	0.7	0.2	.011
1000	150	0.7	0.2	.020
1000	200	0.7	0.2	.011

Table 2. Results for Classical approaches

Board Size	*Backtracking*	*Brute Forcing*
4 X 4	Execution time: 1 msec	Execution time: 15 msec
5X 5	Execution time:1 msec	Execution time: 662 msec
6 X 6	Execution time: 1 msec	Execution time: 149000 msec
7 X 7	Execution time: 2 msec	Solution not found until Time-out
8 X 8	Execution time: 2 msec	Solution not found until Time-out

5 Conclusions and Discussions

In this paper, a new hybrid approach for solving N-Queen Problem has been presented. The approach is based on the combination of ant colony optimization and genetic algorithms. This proposed novel approach will efficiently solve the N-Queen problem. The classical algorithms like Backtracking and Brute-Force are not feasible for CSPs and may or may not give a polynomial time solution. The heuristic approach of Genetic Algorithm was implemented and it gives solution in acceptable amount of time. The performance comparison results of these algorithms are shown below in Table 1 and Table 2.

These heuristic approaches have a lot of scope in future and can be hybridized in new ways to evolve better and efficient solutions for solving real life Constraint Satisfaction Problems. Table 1 given below shows the results for Genetic Algorithm on N

(N=8) Queen problem. Table 2 given below shows the result for classical approaches. As the value of N will exceed 20, the classical approaches will become unfeasible and heuristics will give us the efficient solution.

6 Future Work

Our future work includes working on other CSPs. We plan to work upon the challenging CSP i.e. timetabling problem.

Timetabling is one of the common problems of scheduling which can be described as the allocation of resources for factors under predefined constraints so that it maximizes the possibility of allocation or minimizes the violation of constraints.We aim to apply different heuristics on this problem and propose an efficient solution for it.

References

[1] Khan, S., Bilal, M., Sharif, M., Sajid, M., Baig, R.: Solution of n-Queen Problem Using ACO. IEEE (2009)
[2] Crawford, K.D.: Solving the N-Queens Problem Using Genetic Algorithms. In: Proceedings ACM/SIGAPP Symposium on Applied Computing, Kansas City, pp. 1039–1047 (1992)
[3] Božikovic, M., Golub, M., Budin, L.: Solving n-Queen problem using global parallel genetic algorithm. In: EUROCON, Ljubljana, Slovenia (2003)
[4] Martinjak, I., Golub, M.: Comparison of Heuristic Algorithms for the N-Queen Problem. In: Proceedings of the ITI 2007 29th Int. Conf. on Information Technology Interfaces, June 25-28 (2007)
[5] Thanh, N.D.: Solving Timetabling Problem Using Genetic and Heuristic Algorithms. In: Eighth ACIS International Conference on Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing. IEEE (2007)
[6] Kokash, N.: An introduction to heuristic algorithms. Department of Informatics and Telecommunications
[7] Cormen, T., Leiserson, C., Rivest, R.: Introduction to algorithms. MIT Press (1989)
[8] Horowitz, E., Sahni, S.: Fundamentals of computer algorithms. Computer Science Press Inc., Rockville (1978)
[9] Crawford, K.D.: Solving n-Queen problem using genetic algorithms. Tulsa University
[10] Brailsford, S.C., Potts, C.N., Smith, B.M.: Constraint satisfaction problems: Algorithms and applications. European Journal of Operational Research 119, 557–581 (1998)
[11] Khajehzadeh, M., Taha, M.R., El-Shafie, A., Eslami, M.: A Survey on Meta-Heuristic Global Optimization Algorithms. Research Journal of Applied Sciences, Engineering and Technology (June 25, 2011) ISSN 2040-7467

Testing for Software Security: A Case Study on Static Code Analysis of a File Reader Java Program

Natarajan Meghanathan[1],* and Alexander Roy Geoghegan[2]

[1] Jackson State University
Jackson, MS, USA
natarajan.meghanathan@jsums.edu
[2] L-3 Communications
Greenville, TX, USA

Abstract. The high-level contribution of this paper is to illustrate the use of automated tools to conduct static code analysis of a software program and mitigate the vulnerabilities associated with the program. We present a case study of static code analysis conducted on a File Reader program (developed in Java) using the Source Code Analyzer and Audit Workbench automated tools, developed by Fortify, Inc. Specifically, the following software vulnerabilities are discovered, analyzed and mitigated: (i) Denial of Service, (ii) System Information Leak, (iii) Unreleased Resource (in the context of Streams) and (iv) Path Manipulation. We describe the potential risks in having each of these vulnerabilities in a software program and provide the solutions (including the code snippets in Java) to mitigate these vulnerabilities. The proposed solutions for each of these vulnerabilities are more generic and could be used to correct such vulnerabilities in software developed in any other programming language.

Keywords: Denial of Service, Path Manipulation, Sanitization, Static Code Analysis, Testing for Software Security, Vulnerability.

1 Introduction

Static Code Analysis (also invariably referred to as 'Source Code Analysis') is the process of examining a piece of code without actually executing it [1]. This allows the analyst to see everything that the code does and to consider the program as a whole, rather than just as a sequence of individual lines. Traditionally, static code analysis has been used to evaluate software with respect to functional, semantic and structural issues such as type checking, style checking, program verification, property checking and bug finding. With the growth of the Internet and significant increase in the attacks on software applications, the use of static code analysis to analyze the security aspects of software is gaining prominence. There are a number of issues which must be evaluated when performing a security analysis on a piece of code. These include questions on the security of the basic design of the application, the underlying technologies involved, the potential threats and the risks associated with the threats leading to an attack, the outcome of an attack, and the security controls currently implemented for the

* Corresponding author.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 529–538.
springerlink.com

application. Answering these questions, through a manual analysis of the source code, can prove to be time-consuming for an analyst. Providing the right answers to these issues also requires a comprehensive knowledge of possible exploits and their solutions.

The Fortify Source Code Analyzer, SCA [2], can be used to perform a static code analysis on C/C++ or Java code and can be run in Windows, Linux, or Mac environments. The Fortify SCA can analyze individual files or entire projects. The analyzer follows a set of rules which are included in a rulepack, and users may use generic rulepacks or create their own custom sets of rules. The analyzer can create reports in HTML format for easy viewing, as well as reports which may be viewed with the Audit Workbench utility that is included in the Fortify suite of tools. The Audit Workbench allows users to fine-tune the results of a static code analysis and limit the displayed results to those of interest. The Workbench also includes an editor which will highlight the troublesome-code and allow users to make changes to the code within the application. For each generic issue flagged by the analyzer, the Workbench provides a description of the problem and how it may be averted.

With a generic rulepack, the SCA will identify four categories of issues: (i) semantic, (ii) dataflow, (iii) control flow, and (iv) structural issues. **Semantic issues** include those such as system information leaks, present whenever information specific to the program's internal working may be inadvertently provided to a user. An example of this type of semantic issue is the use of the *printStackTrace*( ) method in Java's Exception class. This method outputs the call stack at the point at which the exception occurred and provides a user with information about the program's structure, including method names. While use of the *printStackTrace*( ) method is handy during debugging, it is not advisable to leave such code in a finished product. **Dataflow issues** include those through which data can cause unwanted effects on a program's execution. For example, if a line of text were accepted from a user, incorporated into an SQL query and executed directly onto the database without first checking the user input for correctness, it may perpetrate an SQL-injection attack [3]. This highlights the need for proper input handling. **Control flow issues** are those related to an improper series of commands. An example of a control flow issue would be if a resource were allocated but never released. **Structural issues** can relate to bugs within the code that do not necessarily affect the performance of an application, but are still not advisable. For example, if a password is hard-coded into a program, anyone who can gain access to the code can also gain access to the password.

While the *Source Analyzer* tool is a command-line tool, the *Audit Workbench* utility offers a graphical user-interface which makes it easy for users to view the results of a static code analysis on a set of source code files and correct the issues raised during the source code analysis. The Audit Workbench's input is a report generated by the Source Analyzer. The Audit Workbench's AuditGuide allows a user to control the types of warnings and issues displayed during an audit. Each type of issue can be either turned on or off according to a user's needs. When an audit's settings have been selected, the Workbench displays the results of the audit. The interface displays a list of the issues that have been flagged and groups these issues according to their severity (hot, warning, or info). The original source code file is also displayed so that a user can immediately access the offending code by selecting an issue in the Issues Panel. For each issue that is shown, the Workbench displays a panel providing background information about the issue and suggestions for resolving the issue. The Workbench can also generate reports in different formats.

2 Case Study on Static Code Analysis of a File Reader Program in Java

The objective of the File Reader program testFileRead.java (original source code in Figure 1) is to read the contents of a text file (name input by the user as a command line argument) on a line-by-line basis and output the lines as read. It is critical to make sure the Java program compiles before using any of the automated tools. A run of the Fortify *Sourceanalyzer* utility on the testFileRead.java program generates (see Figure 2) the following vulnerabilities: 3 medium and 3 low. When we forward the results to an Audit Workbench compatible .fpr file (see Figure 3) and open the log file in Audit Workbench (see Figure 4), we see the 6 vulnerabilities displayed in Figure 4 as Warnings (3) and Info (3). We can turn off the particular vulnerabilities we are not interested to find/ know in our code by clicking "Continue to AuditGuide >>". Alternatively, to know about all possible vulnerabilities that may exist in our code, we can select the "Skip AuditGuide" button. In the rest of the paper, we will discuss how to fix the following Warnings (Vulnerabilities) displayed by the Fortify Audit Workbench tool: (i) Denial of Service, (ii) System Information Leak, (iii) Unreleased Resource: Streams and (iv) Path Manipulation.

```
1    import java.io.*;
2    class testFileRead{
3
5      public static void main(String[ ] args) throws IOException{
6
7      try{
8
9        FileReader fr = new FileReader(args[0]);
10       BufferedReader br = new BufferedReader(fr);
11
12       String line = null;
13
14       while ( (line = br.readLine( ) ) != null){
15         System.out.println(line);
16       }
17
18
19       br.close( );
20       fr.close( );
21
22
23     }// try block
24     catch(IOException ie){
25       ie.printStackTrace( );
26       }
27     }
28   }
```

Fig. 1. Original Java source code for the file reader program

```
C:\res\CCLI-2010\Modules-Meghanathan\Static-Code-Analysis-Examples\Ex1_FileReade
r>sourceanalyzer testFileRead.java


[C:\res\CCLI-2010\Modules-Meghanathan\Static-Code-Analysis-Examples\Ex1_FileRead
er]

[F014B0E28C8E6288784927FC772618FE : low : Denial of Service : semantic ]
testFileRead.java(14) : BufferedReader.readLine()

[EDD1323454D69423D2DD7D4D187D22B7 : medium : System Information Leak : semantic
]
testFileRead.java(25) : Throwable.printStackTrace()

[78FA82368471A9D617111E250114E445 : medium : Path Manipulation : dataflow ]
testFileRead.java(9) :  ->new FileReader(0)
    testFileRead.java(5) :  ->testFileRead.main(0)

[865F144B2D584D3CB7CEDB696F19A416 : medium : Unreleased Resource : Streams : con
trolflow ]
    testFileRead.java(9) : start -> loaded : fr.new FileReader(...)
    testFileRead.java(10) : loaded -> loaded : fr.new BufferedReader(..., fr, ..
.)
    testFileRead.java(14) : loaded -> end_of_scope : #end_scope(fr) (exception t
hrown)

[423D552C35C67B4A8F045E1C079B74FB : low : J2EE Bad Practices : Leftover Debug Co
de : structural ]
    testFileRead.java(5)

[ADBD437811B82372BC593D8FB94B74B6 : low : Poor Logging Practice : Use of a Syste
m Output Stream : structural ]
    testFileRead.java(15)

C:\res\CCLI-2010\Modules-Meghanathan\Static-Code-Analysis-Examples\Ex1_FileReade
r>
```

Fig. 2. Warnings generated for the original source code (testFileRead.java)

```
C:\res\CCLI-2010\Modules-Meghanathan\Static-Code-Analysis-Examples\Ex1_FileReade
r>sourceanalyzer testFileRead.java -f res_testFileRead_org.fpr


C:\res\CCLI-2010\Modules-Meghanathan\Static-Code-Analysis-Examples\Ex1_FileReade
r>auditworkbench
```

Fig. 3. Logging the warnings to an Audit Workbench format .fpr file

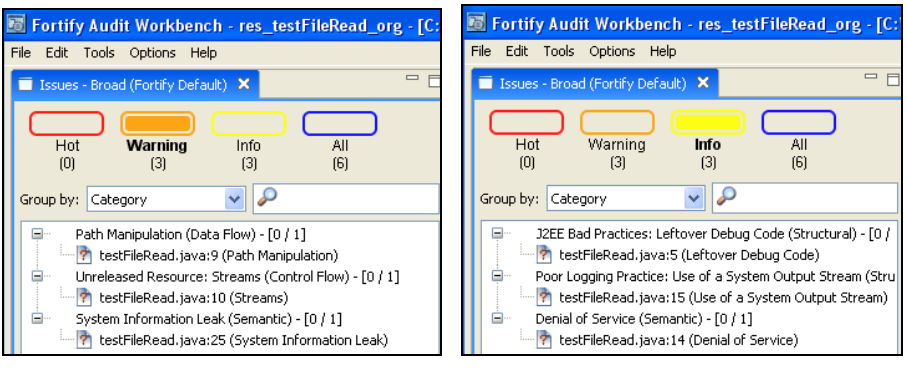


Fig. 4. Vulnerabilities pointed out by Audit Workbench for code in Figure 1

2.1 Denial of Service Vulnerability

A 'Denial of Service' vulnerability is the one using which an attacker can cause the program to crash or make it unavailable to legitimate users [4]. In the Java file reader

program (Figure 1), the *readLine*( ) method invoked on the *BufferedReader* object may be used by the attacker to read an unbounded amount of input. There is no limit on the number of characters that may be buffered and read as a line. An attacker can take advantage of this code to consume a large amount of memory or cause an *OutOfMemoryException* so that the program spends more time performing garbage collection or runs out of memory during some subsequent operation. The solution to remove this vulnerability is to validate the user input to ensure that it will not cause inappropriate resource utilization. In the context of our program, we will limit the number of characters per line that can be read and buffered; it could be 200 or even 1000 characters long (defined through the static variable STR_MAX_LEN in the original testReadFile class); but there needs to be an upper limit to avoid the code from being misused. We create a new *readLine*( ) static method (refer Figure 5) in our testReadFile class and call it from our main program to read every line of the file through the BufferedReader (as displayed in Figure 6).

```
36    public static String readLine(BufferedReader br) throws IOException{
37
38      StringBuffer sb = new StringBuffer( );
39      int intC;
40      intC = br.read( );
41      String line = null;
42      do{
43            if (intC == -1)
44               return null;
45
46       char c = (char) intC;
47
48       if (c = = '\n') {
49         break;
50       }
51       if (sb.length( ) >=  testFileRead.MAX_STR_LEN) {
52         Throw new IOException("input too long");
53       }
54       sb.append(c);
55     } while ( ((intC = br.read( )) != -1) );
56
57      line = sb.toString( );
58
59      return line;
60    }
```

Fig. 5. A newly added readLine( ) static method to the file reader Java program

2.2 System Information Leak Vulnerability and Solution

The "System Information Leak" vulnerability refers to revealing system data or debugging information that may help an adversary to learn about the system and form a plan of attack [5]. In our code, the *printStackTrace*( ) method, called on the object of class 'IOException' of the file reader Java program, has the potential to leak out

sensitive information about the entire application, the operating system it is running under and the amount of care the developers and administrators have put into configuring the program. Depending upon the system configuration, the leaked information may be dumped to a console, written to a log file or exposed to a remote user.

Developers have to take extreme care while deciding what type of error messages should be displayed by their program, even for the purpose of debugging the program to diagnose problems, which may arise for testing pre- and post-release. It is good to turn off detailed error information and preferably include very brief messages, keeping security in mind. For example, even an "Access Denied" message can reveal that a specific file or user exists on the system and it is just the user do not have the requested access to a particular resource [4]. Debugging traces can sometimes appear in non-obvious places such as embedded HTML comments for an error page. To fix the vulnerability in our code, we just print an error message for the particular exception occurred in the *catch* block without printing the entire stack trace.

```
6
7       public static void main(String[ ] args) throws IOException{
8
9        FileReader fr = null;
10       BufferedReader br = null;
11
12       try{
13
14       fr = new FileReader(args[0]);
15       br = new BufferedReader(fr);
16       String line = null;
17
18         while ( (line = readLine(br)) != null){
19                System.out.println(line);
20                line = null;
21         }
22
23      }// try block
24      catch(IOException ie){
25                System.out.println("IOException occurred ");
26      }
27
28      finally{
29        if (br != null)
30                  br.close();
31        if (fr != null)
32                  fr.close();
33      }
```

Fig. 6. Segment of the file reader program with both the Information leak and the Unreleased resource streams vulnerabilities fixed

2.3 Unreleased Resource Streams Vulnerability and Solution

The Unreleased Resource vulnerability occurs when the program is coded in such a way that it can potentially fail to release a system resource [4]. In our file reader program, we have vulnerability with the *File Reader* and the associated *Buffered Reader* streams being unreleased because of some abrupt termination of the program. If there is any exception returned from the *readLine*( ) method, then the control immediately switches from the *try* block to the *catch* block and the two streams 'fr' of *FileReader* and 'br' of *BufferedReader* will never be released until the OS (operating system) explicitly forces the release of these resources upon the termination of the program. From a security standpoint, if an attacker can intentionally trigger a resource leak, the attacker might be able to launch a denial of service attack by depleting the resource pool.

```
63      public static int sanitize(String filename){
64
65        if (filename.indexOf( (int) '/') != -1){
66          System.out.println(" invalid argument... You cannot read from a directory other than the current one");
67              return -1;
68          }
69
70          if ( !filename.endsWith(".txt")){
71              System.out.println(" you can read only a text file with a .txt extension..");
72              return -1;
73          }
74
75          return 0;
76      }
```

Fig. 7. Code for the sanitize( ) method to validate the filename input by the user

A solution to the "Unreleased Resource" vulnerability is to add a *finally* { … } block after the *try* {…} *catch* {…} blocks and release all the resources that were used by the code in the corresponding *try* block. Note that in order to do so, the variables associated with the resources have to be declared outside and before the *try* block so that they can be accessed inside the *finally* block. In our case, we have to declare the *FileReader* object 'fr' and the *BufferedReader* object 'br' outside the *try* block and close them explicitly in the *finally* block. The modified code segment that fixes both the Information Leak and the Unreleased Resource Streams vulnerabilities is shown in Figure 6.

2.4 Path Manipulation Vulnerability and Solution

Path Manipulation vulnerability occurs when a user input is allowed to control paths used in file system operations [5]. This may enable an attacker to access or modify otherwise protected system resources. In our file reader program, the name of the file we would like to read is passed as a command-line argument (args[0]) and we directly

insert this as the parameter for the *FileReader* constructor. Path manipulation vulnerability is very risky and should be preferably avoided in a code. For example, if the program runs with elevated privileges, directly embedding a file name or a path for the file name in our program to access the system resources, could be cleverly exploited by a malicious user who might pass an unexpected value for the argument and the consequences of executing the program with that argument may turn out to be fatal.

```
12      try{
13
14      Scanner sc = new Scanner(System.in);
15      String filename = sc.next( );
16
17      if (sanitize(filename) != -1){
18
19        fr = new FileReader(filename);
20        br = new BufferedReader(fr);
21        String line = null;
22
23          while ( (line = readLine(br)) != null){
24                System.out.println(line);
25                line = null;
26           }
27
28        } // sanitized successfully if block
29
30     }// try block
31
32      catch(IOException ie){
33                System.out.println("IOException occurred ");
34      }
35
36     finally{
37
38       If (br != null)
39                  br.close( );
```

Fig. 8. Code Segment with the Path manipulation vulnerability fixed by obtaining the user input using a Scanner object and sanitizing the input

Some of the solutions to prevent Path manipulation [6] include: (i) Developing a list of valid values the user can enter for the arguments/ variables in question and the user cannot choose anything beyond those values. For example, in the file reader program, we could present the user the list of files that could be read and the user has to select one among them. (ii) Another solution is to have a White list of allowable characters in the user input for the argument/ variable in question. For example, if a

user is allowed to read only a text file, the last four characters of the user input should be ".txt" and nothing else. (iii) Another solution is to have a Black list of characters that are not allowed in the user input for the argument/ variable in question. For example, if the user is not permitted to read a file that is in a directory other than the one in which the file reader program is running, then the input should not have any '/' character to indicate a path for the file to be read. We have implemented solutions (ii) and (iii) through the *sanitize*( ) method, the code of which is illustrated in Figure 7; the method should be called by passing the name of the file input by the user and the rest of the file reader program can proceed only if the *sanitize*( ) method returns a positive result (i.e., 0). Figure 8 illustrates the segment of the file reader program Java code with the Path Manipulation vulnerability fixed by obtaining the user input using a Scanner object and sanitizing the input.

3 Conclusions

Software security is a rapidly growing field and is most sought after in both industry and academics. With the development of automated tools such as Fortify Source Code Analyzer, it becomes more tenable for a software developer to fix, in-house, the vulnerabilities associated with the software prior to its release and reduce the number of patches that need to be applied to the software after its release. This paper presented a case study of a file reader program (developed in Java) on how to analyze the source code of an application using an automated tool, to capture the inherent vulnerabilities present in the code and to mitigate one or more of these vulnerabilities. We discussed the use of an automated tool called the Source Code Analyzer (SCA), developed by Fortify, Inc., and illustrated the use of its command line and graphical user interface (Audit Workbench) options to present and analyze the vulnerabilities identified in a software program. The SCA could be used in a variety of platforms and several object-oriented programming languages. The four vulnerabilities that are specifically discussed in length and mitigated in the case study include the Denial of Service vulnerability, System Information Leak vulnerability, Unreleased Resource (streams) vulnerability and the Path Manipulation vulnerability. Even though our code is written in Java, the solutions proposed and implemented here for each of these four vulnerabilities are more generic and can be appropriately modified and applied in other object-oriented programming languages. More details about this case study can be accessed online at [7]. We have also posted several modules on software security at our website [8].

Acknowledgments. The work leading to this paper has been partly funded through the U. S. National Science Foundation (NSF) CCLI/TUES grant (DUE-0941959) on "Incorporating Systems Security and Software Security in Senior Projects." The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the funding agency.

References

1. Chess, B., West, J.: Secure Programming with Static Analysis, 1st edn. Addison Wesley (2008)
2. https://www.fortify.com/products/hpfssc/source-code-analyzer.html (last accessed: December 20, 2011)
3. Whittaker, J. A.: How to Break Software, 1st edn. Addison-Wesley (2002)
4. Howard, M., Leblanc, D., Viega, J.: 24 Deadly Sins of Software Security: Programming Flaws and How to Fix Them, 1st edn. McGraw-Hill (2009)
5. McGraw, G.: Software Security: Building Security, 1st edn. Addison-Wesley (2006)
6. Graff, M.G., Van Wyk, K.R.: Secure Coding: Principles and Practices, 1st edn. O'Reilly Media (2003)
7. http://www.jsums.edu/cms/tues/docs/Case-Study-Static-Code-Analysis.pdf (last accessed: December 20, 2011)
8. http://www.jsums.edu/cms/tues (last accessed: December 20, 2011)

Vital Signs Data Aggregation and Transmission over Controller Area Network (CAN)

Nadia Ishaque[1], Noveel Azhar[3], Atiya Azmi[2,3], Umm-e-laila[1], and Ammar Abbas[1]

[1] Department of Computer Engineering,
Sir Syed University of Engineering and Technology, Karachi, Pakistan
{786.nadya,ulaila2002,ammarabbas}@gmail.com
[2] Department of Computer Engineering,
Yanbu University College, Yanbu, Saudia Arabia
at.azmi@gmail.com
[3] Department of Electronic Engineering,
Sir Syed University of Engineering and Technology, Karachi, Pakistan
smartnoveel02@hotmail.com

Abstract. An Intensive Care Unit (ICU) consists of patient bedsides which are equipped with a set of sensors to monitor the condition of the patient. These sensors nodes periodically send the patients' vital signs data to the central monitoring system for storing, analysis and emergency treatment requirements. The medical staff uses this data to ascertain the condition of the patient and to provide necessary medications, in case of abnormality. In this paper a design of Controller Area Network (CAN) based ICU monitoring system is presented. CAN provides high speed and reliable transmission of distributed sensor networks data. The key feature in this design is the proposed aggregation scheme which is adopted to effectively utilize the bandwidth and decrease bus load in order to accommodate more number of patients within the existing network.

Keywords: VCAN, Vital signs, Data Aggregation, CAN node, cost effective solution, Patient monitoring.

1 Introduction

Controller Area Network (CAN) protocol is popular in vehicular industry due to its advantageous features like robustness and cost effectiveness. Recent researches and surveys show that CAN has a number of other industrial applications such as military, avionics and medicine. Due to its feasibility in many medical applications, CAN-based systems have been introduced in some hospitals to control operating room components such as lights, tables, cameras, X-ray machines and patient beds [1].

In an Intensive Care Unit (ICU), the vital signs such as Heart Rate (HR), Stolic Blood Pressure (SBP), Distolic Blood Pressure (DBP), Electrocardiogram (EKG),

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 539–548.
springerlink.com

Oxygen Saturation (SPO_2), Temperature (Temp) determine the criticality of patient's condition. These vital signs data from bedside is measured and transmitted to the central monitoring station.

A Controller Area Network based human vital sign data transmission protocol (VCAN) is proposed in [2]. Data aggregation is the process of aggregating the data from multiple sensors to eliminate redundant transmission and provide fused information to the base station [3]. VCAN suggested that the data of vital sign sensors should be aggregated at patient's bedside, into a single packet before transmission to remote monitoring location. VCAN based system provides context data entry of normal vital sign values at patient's bedside and alarm generation incase of exceeding vital sign values, at bedside as well as at remote monitoring station. The simulation results show considerable increase in number of patients that can be accommodated on the network.

This paper, presents the aggregation scheme for CAN-based vital sign monitoring system for effective utilization of bandwidth by minimizing the bus load.

We developed algorithms for transmission node and receiving node of VCAN based vital signs monitoring system. The transmission algorithm performs the acquisition of data form the sensors and aggregates the six vital signs data. The aggregated data is transmitted over CAN bus to remote monitoring station. The proposed receiving algorithm separates the aggregated data and displays it for physician and other medical staff to update them about the condition of patient.

2 Review Research

This section briefly discusses the recent researches related to CAN-based real time patient monitoring system.

In [4] J.A. Zubairi et al presented a scheme that obtain patient medical data over CAN Network in an ambulance. The individual CAN packets of four patients are aggregated and transmitted over CAN bus. The aggregated CAN packet is transformed into IP packets using CAN-UMTS gateway for transmission over UMTS network to efficiently transmit it to the hospital.

Researches showed the feasibility of CAN for the transmission of medical data. There is a research on a "Home Health Information System Based on the CAN Field Bus" [5]. In this work, Gilles Virone et al established an information system for real time health monitoring for home based on CAN. This information system provides remote monitoring of heath status of the elderly persons at home. The smart sensors were linked to the CAN network via single phone pair wire and RF link.

This research suggests a simple sensor data aggregation algorithm for sensors on patient bedside, that periodically transmit patient vital signs to central monitoring station. The patient's biomedical data transmission using CAN protocol provides reliability and error free transmission with good latency.

3 Controller Area Network

The Controller area network (CAN) is a serial communication protocol which provides efficient intercommunication between real time sensor networks. It has a very efficient error detection and message latency time. The bus access mechanism is based on CSMA which help to avoid the collisions when multiple nodes start transmitting data at the same time [6].

CAN protocol defines four different type of frames i.e. data frame, remote frame, error frame, overload frame. The following figure shows the data frame of CAN.

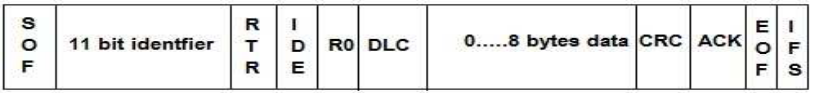


Fig. 1. Data frame of CAN protocol

The data frame contains 0 to 8 bytes of data. The Data Length Code (DLC) is a 4-bit control field which defines the number of bytes present in the data field of data frame.

4 System Descriptions

In the proposed system, a bedside is considered as a CAN node which includes the set of sensors for measuring vital signs of the patient at a bedside. The vital signs data from these sensors is encapsulated in CAN data frame by the CAN chip. The identifier of each CAN message contain the ID of the node from which it is transmitted. This helps the discrimination of data at the central monitoring station. Each node has assigned an ID which is transmitted with the vital sign data. The node ids are assigned in increasing order based on the idea that the condition of any patient may get critical with respect to the other patient in an ICU. In fact, each patient in the ICU is intensively ill. There are finite nodes (patient beds) and the scenario does not require prioritization of any node on the other.

This monitoring data once received at central monitoring station can be stored and further transmitted via intranet or internet for telemedicine purpose.

Figure 2 shows the block diagram of the VCAN system in which several CAN nodes shares a common bus. Each node transmits the human vital sign data from the sensors to the Central Monitoring System. Central monitoring system can also put the data on internet so the monitoring of the patient can be done at remote distance via internet.

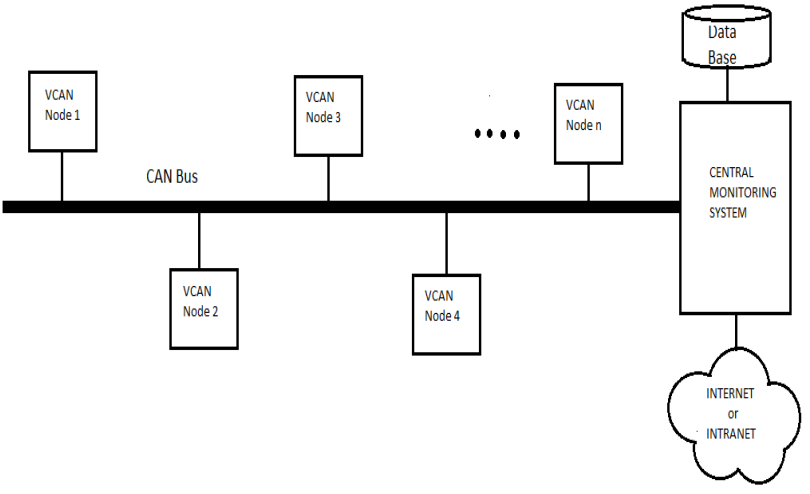


Fig. 2. Block Diagram of VCAN based Monitoring System

4.1 Monitoring Period of Vital Signs

The proposed system is based on the observation that some vital sign parameters do not change significantly in short period of time such as temperature, stolic and diastolic blood pressure, heart beat and SPO_2. Hence, we proposed that data of these sensors be collected periodically after fixed interval of time. However the EKG data will be transmitted continuously at a rate of 300 samples per minute (typical sampling rate of EKG).

Table 1 shows the monitoring period of each parameter in term of the typical sampling rates.

Table 1. Sampling Rate of Vital Signs [2]

Vital Sign	Sampling Rate
Electrocardiogram (EKG)	300 sample/min
Stolic blood pressure (SBP)	1 sample/min
Distolic blood pressure (DBP)	1 sample/min
Heart Rate (HR)	2 sample/min
Temperature (T)	1 sample/min
Oxygen Saturation (SPO_2)	2 sample/min

4.2 Periodic Data Collection and Transmission

Data gathering is defined as the systematic collection of sensed data from multiple sensors to be eventually transmitted to the base station for processing [4]. In our scenario, each bedside unit acts as a CAN node with six monitoring sensors. Each sensor measures one vital sign. All the six sensors are connected to a host processor and are multiplexed so that host processor can collect data from one sensor at a time, as shown in Figure 3.

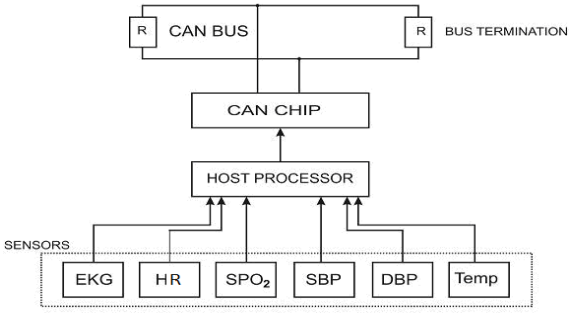


Fig. 3. Transmitting CAN Node

The host processor of the CAN node collects digitized sensor data and aggregates it in a single data frame. During the aggregation the node places the data of each sensor in the proposed order according to Table 2.

Table 2. Arrangement of Vital Sign Data in Data Field of CAN Frame

Byte Number	Vital Sign
1 and 2	EKG
3	HR
4	SPO_2
5	SBP
6	DBP
7	Temp

However, the collection of data from all sensors is not continuous because the fact that not all vital sign change very quickly like Temperature, SBP, DBP. Hence the transmitting node collects and sends data periodically as shown in Fig 3.

According to Fig 4, the CAN node sends EKG data continuously. After every 30 sec it sends aggregated data of HB and SPO_2 along with EKG. Similarly, after 1 minute CAN node encapsulates the data of all six vital signs in a single frame and sends the data on CAN bus.

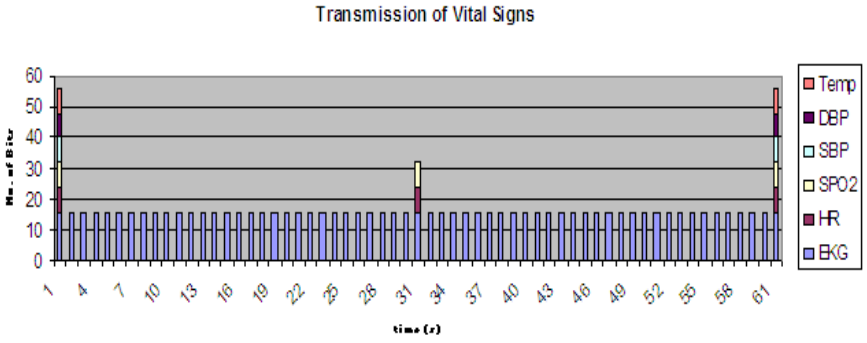


Fig. 4. Transmission of Vital Signs

According to Fig 4, it can be seen that each transmitting node transmits three types of data messages.

1. The first type of data message contains the data of all the vital signs i.e. DLC = 7
2. The second type contains EKG only i.e. DLC= 2
3. The third type contains the data of three vital signs (EKG, HR and SPO_2) i.e. DLC = 4.

4.3 Vital Signs Data Aggregation Algorithm

When the CAN node starts, it aggregates and sends the digitized data of all six sensors in the proposed order mentioned in Table 1. After sending the first message the CAN node intializes two timers T1 and T2 with timeout values 30 and 60 seconds respectively. The CAN node then starts sending the EKG data only to the bus. When timer T1 expires, the CAN node checks for T2 and performs one of the following operations:

- If T2 timeout in not expired the CAN node sends the aggregated data of EKG, HR and SPO_2 and resets T1.
- If T2 timeout is expired the CAN node sends the aggregated data of all the six vitals and resets both timers, T1 and T2, for the next transmission.

The r0 bit in the CAN frame which was reserved for future enhancement is utilized to set the alarm in case if any vital sign parameter indicates the aberrant condition of the patient.

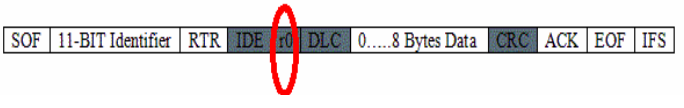


Fig. 5. Standard CAN Frame and R0 bit Highlighted

In the system proposed in [1], the Context Data Input (CDU) unit allows the medical staff or doctors to set the normal range of values for the vital sign parameters according to his age, gender, history and current condition. These values are used by

the CAN node for the comparison between the normal values and the currently measured values of vital signs. If the CAN node detects any vital sign, crossing the normal range it will set the r0 bit to inform the monitoring system so the alarms could be generated. Fig 6 gives Algorithm of sending node at patient's bedside.

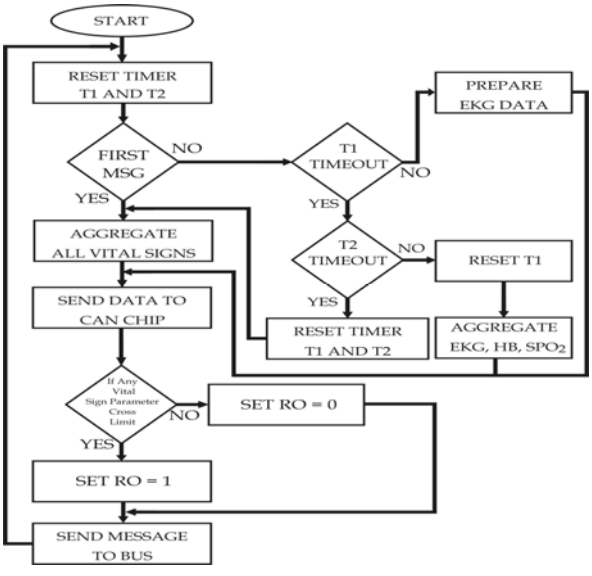


Fig. 6. Algorithm for Sending Node at Patient Bedside

4.4 Data Separation at Receiving Node

At the receiving node, the CAN chip will separate the aggregated data and send respective data to the six host processors. Each host processor is dedicated to receive and display the value of a vital sign at the central monitoring station. The block diagram of receiver CAN node is shown in Fig 7.

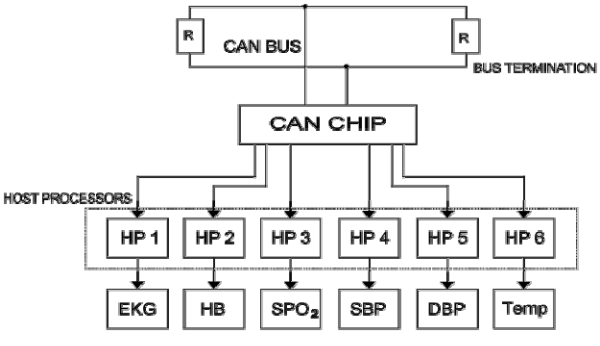


Fig. 7. Receiving CAN Node

The receiving node will disseminate the data among the host processors according to the data length, after checking the DLC field of the message. There could only three type of data frames, with value of DLC be 2, 4 or 7 as described in section 3.

– If the value of DLC is 2 then the message contains 2 bytes EKG data only.
– If DLC is 4, the message contains 2 bytes EKG, 1 byte HR and 1 byte SPO_2 data.
– If DLC is 7, then the data contains data of all vital signs.

Fig 8 shows the algorithm of separation of aggregated data at Central Monitoring Station. The receiving node also checks the status of r0 bit. If this bit is high, the alarm is generated at the Central Monitoring station to inform the paramedic staff about the anomaly in condition of patient.

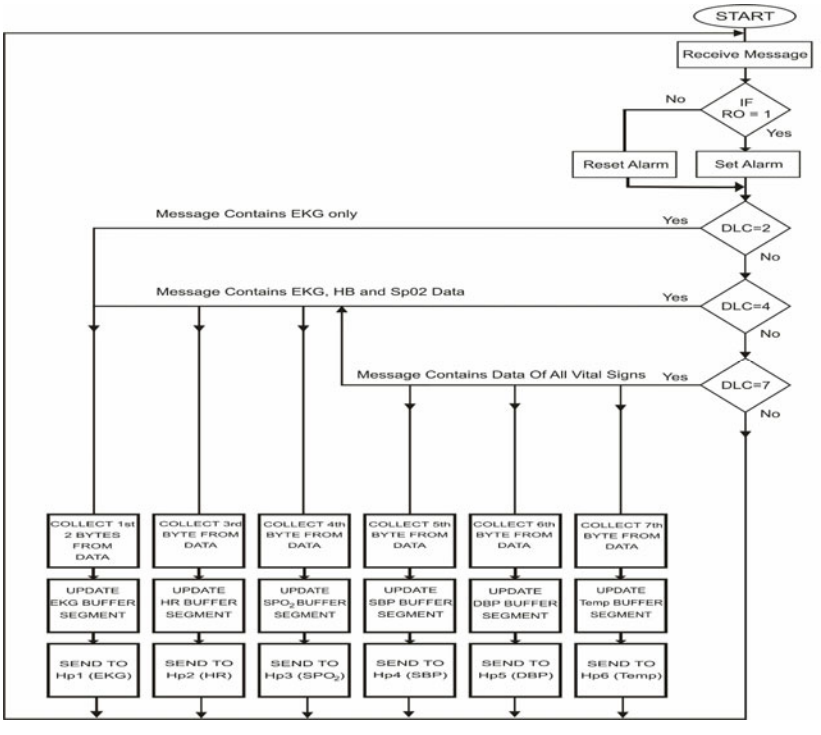


Fig. 8. Algorithm for Receiving Node at Central Monitoring Station

5 Results and Discussion

This section presents that aggregated data transmission reduces the overhead bits.

For example HB, SPO_2, SBP, DBP and temperature are represented by 8 bits each and EKG is represented by 16 bits. For six vital sign data frames sent separately the total number of bits would be 332 bits. Now if the data of all the six vital signs is aggregated to single data field the total data bits will be 56 bits and the total length of CAN frame is 102 bits.

It can be seen that by proposed data aggregation method the bus load can be decreased up to 70%. Fig 9 illustrates the comparison between the aggregated and unaggregated data transmission for three possible sampling rates that are 256, 300 and 500 samples per second. The y-axis shows the number of bits that are required to transmit the data of the six vital signs at different sampling rates.

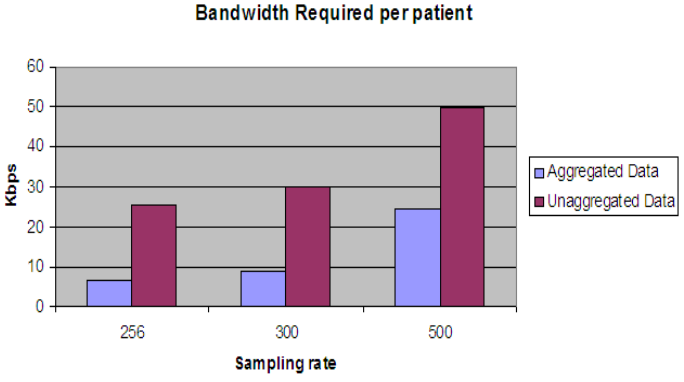


Fig. 9. Aggregated Vs Unaggregated Data Transmission over CAN

The maximum patients that can be accommodated are shown in the Fig 10. The three bars are showing the maximum number of patient for 20% and 50% loss in total bandwidth.

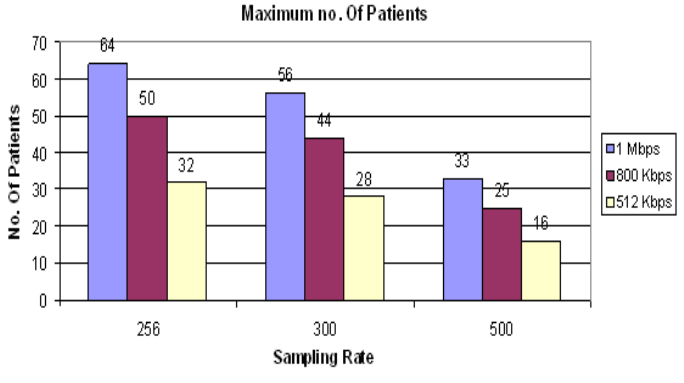


Fig. 10. Maximum Number of Patients Accommodated by System with 1Mbps, 800kbps and 512kbps

Fig 8 depicts number of possible patient's that can be monitored by our proposed system with using proposed VCAN protocol (based on CAN2.0A protocol). In above figures, the three bars are showing the maximum patients that can be accommodated for the available bandwidth. With an ideal condition and availability of maximum bandwidth i.e. 1 Mbps, maximum patient's data that can be transferred to remote

monitoring location are 64, 56 and 33 with bandwidth requirements of 24kbps, 28 kbps and 40kbps depending upon frame rate.

6 Conclusion

In this research work we proposed data aggregation algorithm for CAN based vital sign monitoring system. The results show that the proposed process of aggregating data from multiple sensors eliminate redundant transmission and decrease up to 70% bus load. Moreover, the system can accommodate up to 64 patients even if the highest sampling rate for EKG i.e. 500 samples per second is selected for each patient. On the other hand, this work also proposes a unique and very simple idea of identifying and separating the aggregated data based on the value of Data Length Code Field.

References

1. Controller Area Network (CAN) Overview-National Instrument Developer Zone, http://zone.ni.com/devzone/cda/tut/p/id/2732
2. Azmi, A., Ishaque, N., Abbas, A., Soomro, S.: VCAN-Controller Area Network Based Human Vital Sign Data Transmission Protocol. In: Lin, S., Huang, X. (eds.) CSEE 2011, Part I. CCIS, vol. 214, pp. 290–296. Springer, Heidelberg (2011)
3. Pandey, V., Kaur, A., Chand, N.: A review on data aggregation techniques in wireless sensor network. Journal of Electronic and Electrical Engineering 1(2), 01–08 (2010)
4. Zubairi, J.A.: Aggregation Scheme Implementation for Vital Signs Data over the Network. In: 6th International Conference on High-Capacity Optical Networks and Enabling Technologies (HONET). IEEE Press, Piscataway (2009)
5. Virone, G., Nourey, N., Jean-Pierre, T., Rialle, V., Demongeot, J.: Home Health Information System Based on the CAN Field Bus. In: 5th IFAC Conference on Fieldbus System and Their Application (2003)
6. Bosch CAN specification version 2.0 (2011), https://www.esd.cs.ucr.edu/webres/can20.pdf

A Comparative Study on Different Biometric Modals Using PCA

G. Pranay Kumar, Harendra Kumar Ram, Naushad Ali, and Ritu Tiwari

ABV-Indian Institute of Information Technology and Management, Gwalior
gpk199027@gmail.com, harendra435@gmail.com, naushad56@gmail.com, tiwariritu2@gmail.com

Abstract. Multimodal Biometrics is a rising domain in biometric technology where more than one biometric trait is combined to improve the performance. It is well proved that multimodal biometrics has much higher efficiency than unimodal biometrics. In this research, a comparative study of multimodal biometrics has been done with four biometrics has been considered, three static, one behavioral namely face, ear, palm and gait respectively. The paper mainly focuses on four cases unimodal, bimodal, 3-modal, 4-modal biometrics and using PCA as the base feature extraction method in all cases. The training and testing algorithm has been embedded in PCA itself. the results of all cases are compared and that multimodal cases has much higher efficiency than unimodal and then comparing multimodal cases i.e. bi modal,3- modal,4-modal, there was not much difference and shows that, increasing the number of biometrics doesn't make much difference.

Keywords: Multimodal biometric, PCA, Feature extraction, Face Recognition, Ear Recognition, Palm Recognition, Gait Recognition.

1 Introduction

There are so many biometric systems, but with low recognition rates as they are unimodal. There may be situation that if there is no good image then the rate would be low as it depends on single image and then comes the concept of multimodal biometrics in which two or more than two features are taken and which are well proved with recognition rates better than unimodal systems [2].In unimodal there can be so many methods by taking different features for recognition and in multimodal, different combination of features affect recognition rates. so there is confusion of recognition rates on different modals which may be unimodal or multimodal and which features or combination of features have best recognition rates, these are of one part and feature extraction methods are of second part which effects recognition rates as it is known that there are as many feature extraction methods. In this paper recognition rates of unimodal and multimodal system with four case studies are discussed and used PCA as the base feature extraction method for all cases, four features have been used namely face, palm, gait and ear and test algorithm has been embedded in same PCA algorithm and and then recognition rates of all cases are compared [3].Face recognition is the most common biometric characteristic used by humans to make a

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 549–557.
springerlink.com

personal recognition. The most popular approaches are location of facial attributes eyes, nose and their geometrical dimensions. Palms of the human hands contain unique pattern of ridges and valleys. So in palm print usually variations in ridges or bifurcation in ridges are used as features. Since palm is larger than a finger, palm print is expected to be even more reliable than fingerprint [8]. Palm print scanners have to capture larger area with similar quality so in palm print the cost is more than finger print. For ear, usually the shape of the ear and the structure of the cartilaginous tissue of the pinna are considered as features as they are distinctive [1] [16]. Gait is the peculiar way one walks which differs and it is a complex spatio-temporal biometrics. Gait is a behavioral biometric and may not remain the same over a long period of time, due to change in body weight **or** serious brain damage. The brightness of pixel of biometrics are used as features for recognition in our research.

There have been done so many research on unimodal,bi modal,3-modal etc individually with different feature extraction techniques.In our research comparison between these modals has been done with base feature extraction method PCA and unique testing algorithm for training and testing and important results has been checked out which would help in further research in this area.

2 Methodology

Four biometrics have been taken three static and one behavioral namely face, palm print, ear and gait respectively and four cases have been made with PCA as the feature extraction method. First case with individual performances, second case with bimodal comparisons, thirdly with combining three features and comparing recognition rates and fourth case with combining four features. And at last comparing PCA based recognition rates of all cases [13].

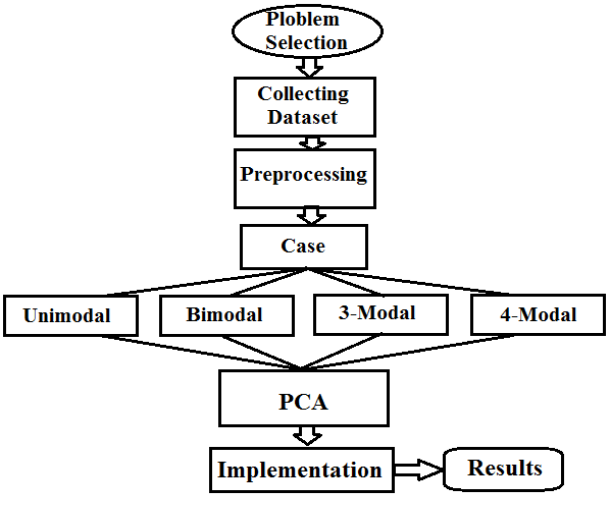


Fig. 1. Pictorial representation of our approach

The database has been taken from ABV-IIITM ,consists of 40 subjects with each class comprises of 5 images of a person and then transforming color image of subjects to 8 bit image (black & white),as in PCA grey image has to be used. After that four cases has been made unimodal,bimodal,3-modal and 4-modal as shown in fig-1,then applying PCA which is our feature extraction method on all cases and then finally training and testing is done with the algorithm and recognition rates of all cases are compared.

2.1 Feature Extraction and Reduction Using PCA

A feature is actually represented by feature vector. The feature vector is an n dimensional vector which carries out measurements on the features [4]. A feature is robust if it will provide consistent results across the entire application domain [5].

The images we analyze are stored in the form of a two- dimensional data array, in which each datum is referred to as a pixel (picture element). We refer an individual pixel located at row i and column j by the notation [6]:

B (i,j) = the brightness of the image at the point (i,j)

For a particular image the each pixel point value considered as feature, and collection of these pixel point represent the feature vector. But due to the large dimension of image feature vector size to large [11]. And also some feature for each image not playing the important role in classification, so there is need to reduce this large dimensionality of feature vector.

Principal component analysis is one method to reduce the dimensionality of variable. It can also be called Karhunen Loeve transform, named after kari Karhunen and michell Loeve[12]. It is a feature extraction and data reduction methods which extract data, remove redundant information, highlighting hidden feature, and show the relationship that exist between observations [7]. A simple approach is here to extracting the information contained in image of face and gait, this extracted information show the variance of image, this information is needed to encode and compare individual face images. Mathematically we wish to find the principal component of the distribution of image, or the covariance matrix of the set of images [5]. Extracted information through principal component analysis is called eigen vector. These Eigen vector are ordered, which show the different amount of variance of images. Through PCA the recognition process can be done by taking K subject images and setting a sampled image S_i. The images are mean centered by subtracting the mean image from each image vector. Then Computing the deviation of each image from mean image by subtracting mean image from each image.Next we are trying to find a set of Eigen value which has the largest projection on to each of wi,which is deviation image. the convergence matrix can be defined as- $C=WW^T$ .Then we calculate first K-1 eigen value and its corresponding eigen vector,$e_{i=}wd_i$, $\lambda_{i=}u_i$ .Then Eigen vectors are normalized.The simplest method for determining which face class provides the best description of an input facial image is to find the face class l that minimizes the Euclidean distance

$$\varepsilon = ||\Omega - \Omega_1|| \quad (1)$$

Where is a vector describing the l[st] face class. If is less than some predefined threshold θ_ε, a face is classified as belonging to the class l.

2.1.1 PCA Used in Different Biometrics

Fig. 2. Grey images of four biometrics

As per fig 2, for face recognition, the main idea of PCA is to express the large 1-D vector of pixels constructed from 2-D facial image into the compact principal components of the feature space. This can be called eigenspace projection. Eigen space is calculated by identifying the eigenvectors of the covariance matrix derived from a set of facial images (vectors) [13].

For Palmprint, each pixel point value considered as feature, and collection of these pixel point represent the feature vector , so we take some features in a image which represent brightness of image at that point pixel the technique of PCA described above is used to extract feature[10][11].

For ear, we use same as palmprint[9], i.e. we use a pixel point and record its brightness at that pixel point, but here we take center of ear and from that point horizontally, vertically and diagonally lines are drawn intersecting and from those lines, each pixel is selected and brightness of each pixel is noted and taken as feature [14].

For Gait we have taken the side view image, and taken features of side view part of gait and PCA is applied here same as face of taking brightness of pixel as feature[17].

2.2 Implementation and Simulation

A total cf 40 samples has been taken i.e.40 individuals are taken where each image is of 600X800 dimensions and considered four biometrics 3 static and 1 behavioural i.e. face, ear, palmprint and gait respectively. We converted all images from RBG to 8-bit (Black&White).We used PCA as the feature extraction method for all cases and recognition rates were compared. Four cases has been taken and compared and database of 40 classes of people are taken, where each class comprises of 5 images of a person, with each image of dimnsion320X240

Case 1 comprises of unimodal biometric i.e. individual biometrics are taken and considered individually. The biometrics are: face, ear, gait, palm. A total of five images of each person are taken in which four images of each person from the database are for training and one for testing. Each image was of dimension 320X240 and an input matrix of 76800X160 is obtained. After giving the input, we calculated the mean image and calculated deviation of each image from mean image. Then surrogate matrix is calculated using covariance matrix which is 160X160.Finally we get eigen vector also known as feature vector matrix is 90X160. Then the Euclidean distance for Eigen

matrix is calculated, which will be a matrix of 1 X N (160) .Then minimum value for all column values is found, if it is within threshold value then column value is returned, that number represents the image number. The number shows that the test image is nearest to that particular image from set of training images, if value is above threshold then not matched. Our PCA algorithm directly gives the recognition rate by dividing total number of matches by total number of test images (40).This is applied for face, ear, palmprint and gait and individual recognition rates of each biometric are produced and results were discussed in section 3[15].

Case 2 comprises of bimodal biometric with total of six combinations where 10 images of each person are taken, four of one biometric and four of other biometric for training and two individual biometrics for testing and then continuing with same process as in case1.

Case 3 comprises of 3- modal biometrics with total of four combinations where total of 15 images of each person are taken, four of one biometric, four of second biometric and four of third biometric for training and one image from each individual biometrics for testing and then same process as in case 1.

Case 4 comprises of four modal biometrics , in which all four biometrics are combined in one and here 20 images of each person are taken ,four of one biometric, four of second biometric, four of third biometric and four of fourth biometric for training and four individual biometrics for testing. While in case of testing the recognition rate can be obtained by –

*No of matched samples/Total no of Samples *100*

As in below section, the in the case of unimodal biometrics, face has the recognition rate 90% which means that approx 36 has matched in testing.After this implementation part recognition rates of all cases are obtained and in next section results obtained are discussed.

3 Results and Discussion

The various results of unimodal, bi modal, 3-modal and 4- modal are given in section3.1 and discussed in section 3.2.

3.1 Results

Case1

Table 3.1. Unimodal Recognition Rates

Sl. No.	Biometrics Used	Recognition Rate (%)
1	FACE	90
2	EAR	85
3	GAIT	70
4	PALM	92.5

Case 2

Table 3.2. Bimodal Recognition Rates

Sl. No.	Biometrics Used	Recognition Rate (%)
1	FACE+EAR	97.5
2	FACE+GAIT	95
3	FACE+PALM	95
4	EAR+GAIT	90
5	EAR+PALM	97.5
6	PALM+GAIT	87.5

Case 3

Table 3.3. 3-modal Recognition Rates

Sl. No.	Biometrics Used	Recognition Rate (%)
1	FACE+EAR+GAIT	97.5
2	EAR+GAIT+PALM	95
3	FACE+EAR+PALM	97.5
4	PALM+FACE+GAIT	95

Case 4

Table 3.4. 4-modal Recognition Rates

Sl. No.	Biometrics Used	Recognition Rate (%)
1	FACE+EAR+GAIT+PALM	97.5

3.2 Discussion

Case 1

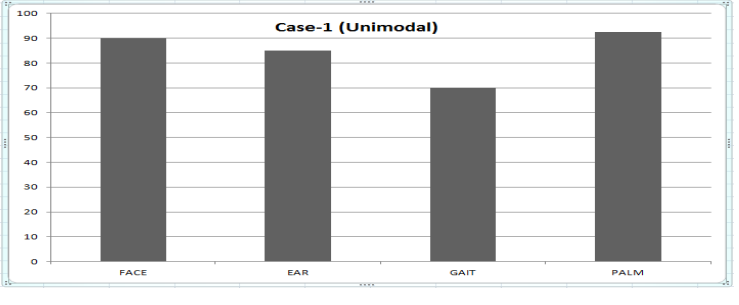


Fig. 3.1. Recognition rate of Unimodal (in percentage)

In case 1, as per figure face has the efficiency of 90% as it is apparent with other images, ear with 85%, gait with 70% and the palm with 92.5%.So here it is noticed that in unimodal the efficiency rate is in the range of 70% -93% by using PCA. Face and palm has higher efficiency than other biometrics,which represents that as they are static, they get more efficiency rate ,but ear has much less recognition rate as ear has features which are not much distinctive and behavioral biometric gait has least recognition rate.

Case 2

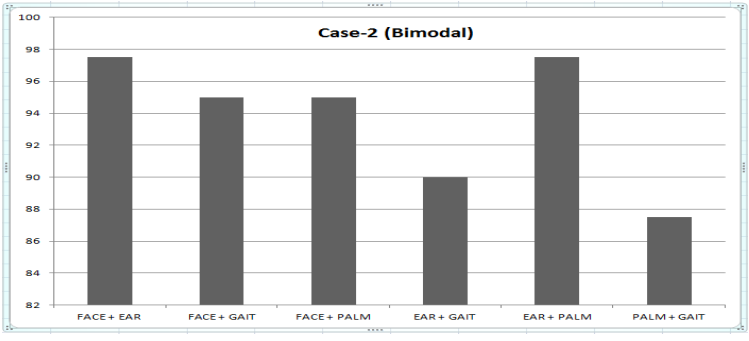


Fig. 3.2. Recognition rate of Bimodal (in percentage)

In case 2 which is bimodal, there are total of six combinations with face+ear has efficiency of 97.5,face+gait has 95% efficiency, face+palm has 95% efficiency, ear+gait has efficiency 90%,ear+palm has efficiency 97.5, Palm+gait has efficiency 87.5%,so by observation it is noticed that in this case the efficiency range has increased from 85%-98

Case 3

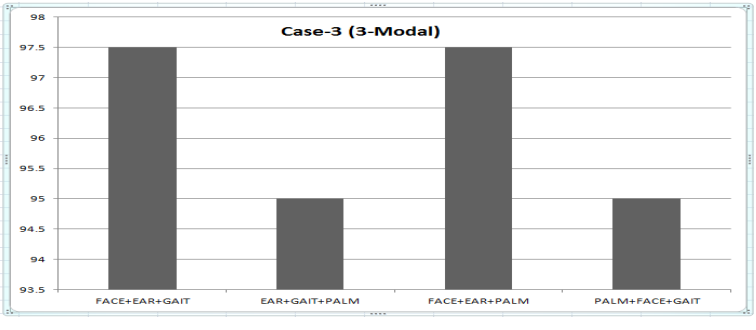


Fig. 3.3. Recognition rate of 3-modal (in percentage)

In case 3 which is 3-modal, there are four combinations with face+ear+gait has efficiency of 97.5, face+ear+palm has 97.5% efficiency, face+palm+gait has 95% efficiency, ear+gait+palm has efficiency 95%.So here as no of biometrics are increased,

the recognition rate has increased compared to previous modals and range of efficiency increased from 95%-98%.

Case 4
In case 4, there is only one combination i.e. face+ear+gait+palm with efficiency 97.5% which means that all test images has matched except one. Here even if two biometrics fails due to failure of camera to take image properly, the system succeeds. This case may have good efficiency but it is not very practical to use 4-modal as this would be costly and burdensome i.e. not much user-friendly process as person may get tensed for giving these many image of him.

4 Conclusions

The 4-modal biometric has high efficiency rate than other modals, 3-modal has higher efficiency than unimodal & bimodal and bimodal has high efficiency than unimodal. As numbers of biometrics are increased, the efficiency rates also get increased. So as numbers of biometrics are increased, not much change is observed compared with unimodal recognition rates. The efficiency rate is not linear with number of biometrics used for recognition.

References

[1] Jain, A.K., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Transactions on Circuits and Systems for Video Technology 14(1), 4–20 (2004)
[2] Jain, A.K., Bolle, R., Pankanti, S. (eds.): Biometrics: Personal Identification in Networked Society. Kluwer Academic Publishers (1999)
[3] Yazdanpanah, A.P., Faez, K., Amirfattahi, R.: Multimodal biometric system using face, ear and gait biometrics. In: 2010 10th International Conference on Information Sciences Signal Processing and their Applications (ISSPA), May 10-13, pp. 251–254 (2010)
[4] Kim, K.: Face Recognition using Principle Component Analysis. In: International Conference on Computer Vision and Pattern Recognition, pp. 586–591 (1996)
[5] Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 1991, June 3-6, pp. 586–591 (1991)
[6] Umbaugh, S.E., Wei, Y.-S., Zuke, M.: Feature extraction in image analysis. A program for facilitating data reduction in medical image classification. IEEE Engineering in Medicine and Biology Magazine 16(4), 62–73 (1997)
[7] Zhao, G., Liu, J.: Application of Principal Component Analysis and Neural Network on the Information System Evaluation. In: Pacific-Asia Conference on Circuits, Communications and Systems, PACCS 1909, May 16-17, pp. 785–788 (2009)
[8] Delac, K., Grgic, M.: A survey of biometric recognition methods. In: 46th International Symposium Electronics in Marine, ELMAR 2004, June 16-18 (2004)
[9] Ahmad, M.I., Woo, W.L., Dlay, S.S.: Multimodal biometric fusion at feature level: Face and palmprint. In: 2010 7th International Symposium on Communication Systems Networks and Digital Signal Processing (CSNDSP), July 21-23, pp. 801–805 (2010)

[10] Adhinagara, Y., Tjokorda Agung, B.W., Retno, N.D.: Implementation of multimodal biometrics recognition system combined palm print and palm geometry features. In: 2011 International Conference on Electrical Engineering and Informatics (ICEEI), July 17-19, pp. 1–5 (2011)

[11] Meraoumia, A., Chitroub, S., Bouridane, A.: Fusion of multispectral palmprint images for automatic person identification. In: 2011 Saudi International on Electronics, Communications and Photonics Conference (SIECPC), April 24-26, pp. 1–6 (2011)

[12] Guo, Z., Zhang, L., Zhang, D.: Feature Band Selection for Multispectral Palmprint Recognition. In: 2010 20th International Conference on Pattern Recognition (ICPR), August 23-26, pp. 1136–1139 (2010)

[13] Bozorgtabar, B., Noorian, F., Rad, G.A.R.: Comparison of different PCA based Face Recognition algorithms using Genetic Programming. In: 2010 5th International Symposium on Telecommunications (IST), December 4-6, pp. 801–805 (2010)

[14] Abaza, A., Ross, A.: Towards understanding the symmetry of human ears: A biometric perspective. In: 2010 Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS), September 27-29, pp. 1–7 (2010)

[15] Cadavid, S., Mahoor, M.H., Abdel-Mottaleb, M.: Multimodal biometric modeling and recognition of the human face and ear. In: 2009 IEEE International Workshop on Safety, Security & Rescue Robotics (SSRR), November 3-6, pp. 1–6 (2009)

[16] Connie, T., Teoh, A., Goh, M., Ngo, D.: Palmprint Recognition with PCA and ICA (November), http://sprg.massey.ac.nz/ivcnz/Proceedings/IVCNZ_41.pdf

[17] Xu, S.-L., Zhang, Q.-J.: Gait Recognition Using Fuzzy Principal Component Analysis. In: 2010 2nd International Conference on E-Business and Information System Security (EBISS), May 22-23, pp. 1–4 (2010)

Methodology for Automatic Bacterial Colony Counter

Surbhi Gupta[1], Priyanka Kamboj[1], and Sumit Kaushik[2]

[1] Seth Jai Parkash Mukand Lal Institute & Technology (JMIT), Radaur
er_surbhi123@gmail.com
[2] Guru Nanak Institutions (GNI), Mullana (Ambala)

Abstract. An increased area of focus in Microbiology is the automation of counting methods. These obstacles include: How to handle confluent growth or growth of colonies that touch or overlap other colonies, How to identify each colony as a unit in spite of differing shapes, sizes, textures, colors, light intensities, etc. Counting of bacterial colonies is complex task for microbiologist. Further in an Industry thousands of such samples are formed per day and colonies on each sample are counted manually, then this becomes a time consuming hectic and error prone job. We proposed a method to count these colonies to save time with accurate results and fast delivery to customers. This proposed research work will count the colonies after 6 to 8 hours priori, saving a lot more time and this work will more efficient because market range for this is about 10,000 only as compare to prior systems.

1 Introduction

Bacterial colony in simple words is a group or cluster of bacteria derived from one common bacteria. Many biological procedures depend on an accurate count of the bacterial colonies and other organisms. The enumeration of such colonies is a slow, tedious task. When counts are made by more than one technician, wide variations are often noted.

To a large extent, accurate colony counting depends on the ability to "see" colonies distinctly, whether viewed by the naked eye or by an automated instrument. Colony morphology is largely a result of the characteristics of the growth media and other environmental conditions. To enhance visibility of colonies and enhance the counting accuracy in an even broader range of applications, it is good practice to employ those procedures that form colonies that are counted easily by their improved size, shape, distribution and contrast.

The counting of bacterial colony is usually performed by well-trained technicians manually. However, this manual counting process has a very low throughput, and is time consuming and labor intensive in practice. To provide consistent and accurate results and improve the throughput, the existing colony counter devices and software were then developed and commercialized in the market. On the other hand, big laboratories may have extremely large counting needs to be accommodated with few automatic counters. Thus, colony counting is a significant budgetary and technical hurdle for laboratories of all sizes.

In this paper, we propose a fully automatic colony counter and compare its performance with manual counting of bacterial colonies. Our proposed method can

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 559–565.
springerlink.com

significantly reduce the manual labor by automatically detecting the colonies and count of those colonies efficiently. Bacterial colony counting is tedious and laborious work because these colonies are not easily seen by naked eyes. To count these bacterial colonies manually is very hectic and time consuming process because Bacteria's are grown onto filter for 24 to 48 hours.

To count these bacterial colonies microbiologist uses some dyes so that bacterial colonies appear as colored spots and our problem is to count the number of these bacterial colonies. Further in an Industry thousands of such samples are formed per day and colonies on each sample are counted manually, then this becomes a time consuming, hectic and error prone job.

Goal is to develop software to save time with accurate results and fast delivery to customers. There are so many devices to count these bacterial colonies but these devices ranges about 50,000 to 70,000 according to the Indian currency, that's why these devices are not so much efficient for daily use. This proposed research work will count the colonies after 6 to 8 hours priori, saving a lot more time and this work will more efficient because market range for this is about 10,000 only as compare to prior systems.

Intense Testing is required before actual installation, on different images of filters of types:

- Images in which size & shape of bacterial colonies vary.
- Images containing very dense bacterial colonies.
- Images containing different types of bacterial colonies on same filter.

There are lots of sample of bacteria for which the proposed method will efficiently work. Some of the samples images are:

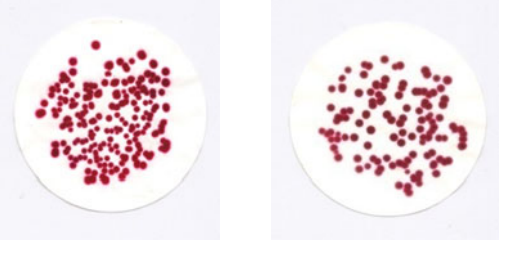


Fig. 1. Sample Input Images

There are various devices available in market to count these colonies but those devices are very costly. We can design an automated bacterial colony counter which used many image processing algorithms such as grayscaling, thresholding, filtering etc. to count these colonies efficiently.

2 Proposed System

Bacterial colony counting is tedious and laborious work because these colonies are not easily seen by naked eyes. To count these bacterial colonies manually is very

hectic and time consuming process because Bacteria's are grown onto filter for 24 to 48 hours. To count these bacterial colonies microbiologist uses some dyes so that bacterial colonies appear as colored spots and our problem is to count the number of these bacterial colonies.

Problem of counting the total number of bacterial colonies present in a sample (filter) have following issues to handle:

- Number of non-overlapping colonies.
- Number of over-lapping colonies.
- Number of edge touching colonies.
- To subtract the count due to noise.
- Colonies of differ-ent size shape and colors.
- And the total count will be the sum of the above five.

2.1 Block Diagram

To count the bacterial colonies, the block diagram for proposed method is given below:

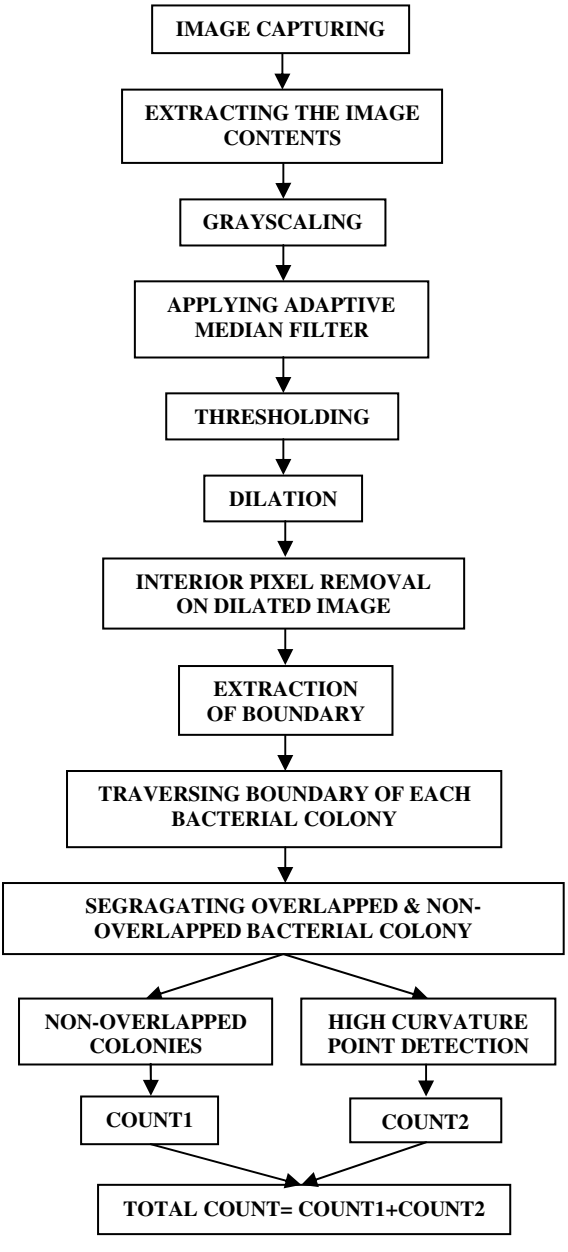


Fig. 2. Block Diagram of Proposed Method

Image Capturing

Bacterial Colonies are grown onto filter for 24 to 48 hours. Some colored dyes are spread over each filter so that bacterial colonies appear as colored spots. Now this filter is kept on a Petri plate. Background is made of black or white intensity so, it

becomes easier to separate the filter from it's surroundings while processing the image. Petri plate is kept in a box containing a digital camera and light arrangement. Images are then captured using this arrangement. The collected images are digitized on a computer utilizing image processing software package that has programming capabilities (note: the system works with any of the software packages with these capabilities). The digitized picture is processed using the various procedures described to separate and detect the colonies present.

Extracting the Image Content
Information about the image will extract in a single dimensional array using pixel grabber function in JAVA.

Gray Scaling
Brightness of the pixels will computed using the NTSC(National Television Standards Committee) color – to – brightness conversion factor.

Thresholding
Thresholding can be defined as mapping of the gray scale into the binary set {0, 1} that is thresholding essentially involves turning a color or grayscale image into a 1-bit binary image. Thresholding algorithm will be applied to the gray scaled image.

Applying Filter
To remove unwanted noise we will use the adaptive median filter which is used for many noises.

Boundary Extraction
Boundaries are linked edges that characterize the shape of an object. They are useful in computation of geometry features such as size or orientation. To extract the boundaries we will use the Dilation and Interior pixel removal method.

Boundary Traversal
Complete image is scanned. It returns the image array containing coordinates of boundary of object and size of image array is equal to the number of points in the boundary

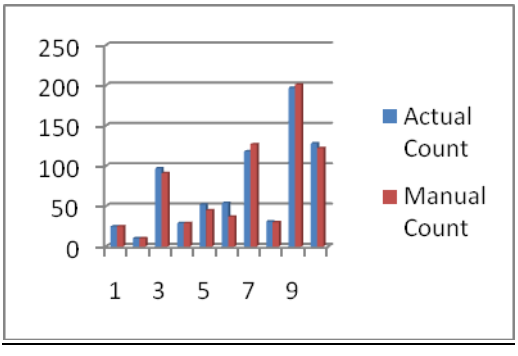
Counting
Colonies will be count according to the results given by above process for overlapping and non-overlapping bacterial colonies.

3 Result

Out of 50 samples on which algorithm were tested, following are 10 samples with variations in contrast, density, color and noise, following output is occurs:

Table 1. Actual Count & Manual Count

Image	Actual Count	Manual Count
1	25	25
2	10	10
3	97	91
4	29	29
5	52	45
6	54	37
7	118	127
8	31	30
9	197	201
10	128	122



Graph 1. Actual Count versus Manual Count

4 Conclusion

Bacterial colony in simple words is a group or cluster of bacteria derived from one common bacteria. We can design an automated bacterial colony counter which used many image processing algorithms such as grayscaling, thresholding, filtering etc. to count these colonies efficiently.

Hence, Images with high contrast and low or medium density give accurate or near to accurate count (99-100%). Images with low contrast or high density give less accurate count (95-98%).

The reason thereof is that in case of low contrast after thresholding shape of colony/colonies gets distorted which leads to appearance of high curvature points along the boundary. These high curvature points (corners) get accumulated in count result.

5 Future Enhancement

Bacterial colony counter will be enhanced to:

- Process the most complex samples and give accurate count.
- Work on any type of samples i.e. samples with very low contrast.
- Tackle any shape and size of colonies.
- Handle all types of noises.
- Detect high curvature points even in very dense overlapping colonies.

Take the shape of a product accepting different samples and giving count of total number of colonies present, which can be used in various biotech and other areas/ industries.

References

1. Shen, W.-Z., Wu, Y.-C., Zhao, L., Zheng, H.: Experimental Study for Automatic Colony Counting System Based on Image Processing. In: 2010 International Conference on Computer Application and System Modeling, ICCASM 2010 (2010)
2. Buzalewicz, I., Wysocka-Król, K., Podbielska, H.: Image processing guided analysis for estimation of bacteria colonies number by means of optical transforms (received April 2, 2010; revised May 8, 2010; accepted May 11, 2010; published June 2, 2010, June 7, 2010)
3. Chen, W.-B., Zhang, C.: An automated bacterial colony counting and classification system, LLC (2009) (published online: February 18, 2009)
4. Men, H., Wu, Y., Li, X., Kou, Z., Yang, S.: Counting Method of Heterotrophic Bacteria Based on Image Processing. IEEE (2008)
5. Zhang, C., Chen, W.-B., Liu, W.-L., Chen, C.-B.: An Automated Bacterial Colony Counting System. IEEE (2008)
6. Chen, W.-B., Zhang, C.: Bacteria Colony Enumeration and Classification for Clonogenic Assay. In: Tenth IEEE International Symposium on Multimedia. IEEE (2008)
7. Trattner, S., Greenspan, H. Member, IEEE, Tepper, G., Abboud, S.: Automatic Identification of Bacterial Types Using Statistical Imaging Methods. IEEE (2004)
8. Zhang, C., Chen, W.-B.: An Effective and Robust Method for Automatic Bacterial Colony Enumeration. In: IEEE International Conference on Semantic Computing (2007)
9. Mosaliganti, K., Chen, J., Janoos, F., Machiraju, R., Xia, W., Xu, X., Huang, K.: Automated Quantification of Colony Growth in Clonogenic Assays
10. Barber, P.R., Vojnovic, B., Kelly, J., Mayes, C.R., Boulton, P., Woodcock, M., Joiner, M.C.: An Automated Colony Counter Utilising A Compact Hough Transform

11. Terada, K., Yoshida, D., Oe, S., Yamaguchi, J.: A Method of Counting the Passing Peopleby Using the Stereo Images. IEEE (1999)
12. Mahapatra, A.K., Harris, D., Nguyen, C.N., Kannan, G.: Evaluation of an IUL Flash & Go Automated Colony Counter. Agricultural Engineering International: the CIGR Ejournal. Manuscript 1368 XI (October 2009)
13. Dahle, J., Kakar, M., Steen, H.B., Kaalhus, O.: Automated Counting of Mammalian Cell Colonies by Means of a Flat Bed Scanner and Image Processing
14. Putman, M., Burton, R., Nahm, M.H.: Simplified method to automatically count bacterial colony forming unit (June 2005)
15. Schier, J., Kovă, B.: Automated Counting of Yeast Colonies Usingthe Fast Radial Transform Algorithm

Sorting of Decision Making Units in Data Envelopment Analysis with Intuitionistic Fuzzy Weighted Entropy

Neeraj Gandotra, Rakesh Kumar Bajaj, and Nitin Gupta

Department of Mathematics, Jaypee University of Information Technology, Waknaghat, Solan (H.P.) India
neeraj.juit@gmail.com, rakesh.bajaj@gmail.com, nitinstat@gmail.com

Abstract. Analysis method used for measuring and evaluating the efficiency of decision making units is basically a linear programming based technique. It has number of inputs and outputs included in the analysis and takes account of the relationship between inputs and outputs. Analysis method has clear advantages over competing approaches such as data envelopment analysis (DEA). In the present paper, we propose a new algorithm for decision making units in context of intuitionistic fuzzy weighted entropy in order to rank decision making units in data envelopment analysis.

1 Introduction

In the fuzzy environment, Data Envelopment Analysis (DEA) was first introduced in [9] as a linear programming based technique used for measuring and evaluating the relative performance of activities in organizations e.g. hospital, bank etc., where the presence of multiple inputs generate multiple outputs. This makes the comparison complex and difficult. DEA is a useful management tool to the assessment and evaluation of decision making units. DEA defines the relative efficiency of decision making unit and has clear advantages over competing approaches. DEA involves identification of units and uses this information to construct efficiency frontiers over the data of available organization units. In [8], the fuzzy efficiency scores of decision making units (DMUs) are counted and maximum entropy as a special class weighting function is used to rank DMUs. Decision making process is very important for functions such as investments, new product development, delivery personnel selections, allocation of resources and many others. In [1] and [2], the concept of intuitionistic fuzzy set (IFS) which is generalization of the theory of fuzzy set has been introduced. [6] proposed multicriteria decision making methods with IFS using various linear programming approaches to generate optimal weights.

A intuitionistic fuzzy set is characterized by two functions - the degree of membership function and a non membership function. It may be noted that the sum of membership function and non membership function must be smaller than or equal to one. The theory of IFS is well suited in dealing with imprecise or uncertain decision information, image edge detection, uncertainty, incompleteness and vagueness in decision making. It has been used to build soft decision making models that can accommodate imprecise information and analyze the extent of agreement in a group of experts. Feasibility

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 567–576.
springerlink.com

and effectiveness of IFSs are illustrated in its applications of decision making by many researchers such as in [3], [7], [10], [13] and [14].

Definition 1. Atanassov's intuitionistic fuzzy set (IFS) over a finite non empty fixed set X, is a set $\widetilde{A} = \{< x, \mu_{\tilde{A}}(x), \gamma_{\tilde{A}}(x) > \mid x \in X\}$ which assigns to each element $x \in X$ to the set $\tilde{A}$, which is subset of X having the degree of membership $\mu_{\tilde{A}}(x) : X \to [0,1]$ and degree of non-membership $\gamma_{\tilde{A}}(x) : X \to [0,1]$, satisfying $0 \leq \mu_{\tilde{A}}(x) + \gamma_{\tilde{A}}(x) \leq 1$, for all $x \in X$. For each IFS in X, a hesitation margin $\pi_{\tilde{A}}(x)$, which is the intuitionistic fuzzy index of element x in the IFS $\widetilde{A}$, defined by $\pi_A(x) = 1 - \mu_A(x) - \gamma_A(x)$, denotes a measure of non-determinacy.

Definition 2. Let $\widetilde{a_i} = (\mu_i, \gamma_i),\ i = 1, 2, ..., n$, be a collection of intuitionistic fuzzy values, the intuitionistic fuzzy weighted averaging operator is defined as

$$IFWA_w(\widetilde{a_1}, \widetilde{a_2}, ..., \widetilde{a_n}) = \sum_{i=1}^{n} w_i \widetilde{a_i} = \left(1 - \prod_{i=1}^{n} (1-\mu_i)^{w_i}, \prod_{i=1}^{n} \gamma_i^{w_i}\right);$$

where w_i is the weight of $\widetilde{a_i}$, $w_i \in [0,1]$ and $\sum_{i=1}^{n} w_i = 1$.

Definition 3. Let $\widetilde{a_i} = (\mu_i, \gamma_i),\ i = 1, 2, ..., n$, be a collection of intuitionistic fuzzy values, the intuitionistic fuzzy weighted geometric operator is defined as

$$IFWG_w(\widetilde{a_1}, \widetilde{a_2}, ..., \widetilde{a_n}) = \sum_{i=1}^{n} \widetilde{a_i}^{w_i} = \left(\prod_{i=1}^{n} \mu_i^{w_i},\ 1 - \prod_{i=1}^{n} (1-\gamma_i)^{w_i}\right);$$

where w_i is the weight of $\widetilde{a_i}$, $w_i \in [0,1]$, and $\sum_{i=1}^{n} w_i = 1$.

Definition 4. Let $\widetilde{a} = (\mu, \gamma)$ be an intuitionistic fuzzy value, the score of $\widetilde{a}$ is defined by $s(\widetilde{a}) = \mu - \gamma$, s is called score function. The degree of accuracy of $\widetilde{a}$ is defined by $p(\widetilde{a}) = \mu + \gamma$, p is called accuracy function.

Let $\widetilde{a_1} = (\mu_1, \gamma_1)$, $\widetilde{a_2} = (\mu_2, \gamma_2)$ be two intuitionistic fuzzy values,

- If $s(\widetilde{a_1}) < s(\widetilde{a_2})$, then $\widetilde{a_1} < \widetilde{a_2}$;
- If $s(\widetilde{a_1}) = s(\widetilde{a_2})$, then
  (i) $p(\widetilde{a_1}) < p(\widetilde{a_2}) \Rightarrow \widetilde{a_1} < \widetilde{a_2}$;
  (ii) $p(\widetilde{a_1}) = p(\widetilde{a_2}) \Rightarrow \widetilde{a_1} = \widetilde{a_2}$.

In section 2, we have presented and studied the fuzzy CCR Data Envelopment Analysis Model. A brief discussion on Intuitionistic Fuzzy Entropy and weighted entropy in subsections 2.1 and 2.2, respectively, has been given. Further, we have proposed a new algorithm for decision making units in context of intuitionistic fuzzy weighted entropy in order to rank decision making units in data envelopment analysis in section 3. In section 4, we provide illustrative examples to show the validity of the proposed algorithm. Finally, we conclude the paper in section 5.

2 The Fuzzy CCR DEA Model

Let us consider n decision making units and each requires varying amounts of m different fuzzy inputs to produce s different fuzzy outputs. The input oriented fuzzy CCR (Charnes, Cooper and Rhodes) model is in [4] and given by

$$\max \quad E_0 = \sum_{r=1}^{s} u_r \widetilde{O_{ro}}$$

such that

$$\begin{aligned} \sum_{i=1}^{m} v_i \widetilde{I_{io}} &= \widetilde{1}; \\ \sum_{r=1}^{s} u_r \widetilde{O_{rj}} - \sum_{i=1}^{m} v_i \widetilde{I_{ij}} &\leq 0; \quad j = 1, \ldots., n, \\ u_r, v_i &\geq 0; \quad r = 1,2,\ldots,s \quad \text{and} \quad i = 1,2,\ldots,m, \end{aligned}$$

where $\widetilde{I_{io}}$; $i = 1,2,\ldots,m$ and $\widetilde{O_{ro}}$; $r = 1,2,\ldots,s$, are input and output values for DMU_o, the decision making unit under consideration.

The α-cuts of $\widetilde{I_{ij}}$ and $\widetilde{O_{rj}}$ are defined as

$$\begin{aligned} (\widetilde{I_{ij}})_\alpha &= \left(x \in X \,|\mu_{I_{ij}(x)} \geq \alpha\right) = [I_{ij}^l, I_{ij}^u] \\ \text{and} \quad (\widetilde{O_{rj}})_\alpha &= \left(x \in X \,|\mu_{O_{rj}(x)} \geq \alpha\right) = [O_{rj}^l, O_{rj}^u]. \end{aligned}$$

On applying the α-level of fuzzy data envelopment analysis, the following model is formed:

$$\max \quad E_0 = \sum_{r=1}^{s} u_r \left[O_{ro}^l, O_{ro}^u\right]$$

such that

$$\begin{aligned} \sum_{i=1}^{m} v_i \left[I_{io}^l, I_{io}^u\right] &= \widetilde{1}; \\ \sum_{r=1}^{s} u_r \left[O_{rj}^l, O_{rj}^u\right] - \sum_{i=1}^{m} v_i \left[I_{ij}^l, I_{ij}^u\right] &\leq 0; \quad j = 1, \ldots., n, \\ u_r, v_i \geq 0;\ r = 1,2,\ldots,s \quad &\text{and} \quad i = 1,2,\ldots,m. \end{aligned}$$

For measuring the lower and upper bounds of the best relative efficiency of each decision making units with interval input and output data, the following DEA model is achieved:

$$\max \quad (E_0)_\alpha^u = \sum_{r=1}^{s} u_r (O_{ro})_\alpha^u$$

such that

$$\sum_{i=1}^{m} v_i (I_{io})^l_\alpha = \widetilde{1}$$
$$\sum_{r=1}^{s} u_r (O_{ro})^u_\alpha - \sum_{i=1}^{m} v_i (I_{io})^l_\alpha \leq 0;$$
$$\sum_{r=1}^{s} u_r (O_{rj})^l_\alpha - \sum_{i=1}^{m} v_i (I_{ij})^u_\alpha \leq 0; \quad j = 1, \ldots., n, \quad j \neq 0;$$
$$u_r, v_i \geq 0; \quad r = 1,2,\ldots,s \quad \text{and} \quad i = 1,2,\ldots,m.$$

Also,

$$\max(E_0)^l_\alpha = \sum_{r=1}^{s} u_r (O_{ro})^l_\alpha$$

such that

$$\sum_{i=1}^{m} v_i (I_{io})^u_\alpha = \widetilde{1};$$
$$\sum_{r=1}^{s} u_r (O_{ro})^l_\alpha - \sum_{i=1}^{m} v_i (I_{io})^u_\alpha \leq 0;$$
$$\sum_{r=1}^{s} u_r (O_{rj})^u_\alpha - \sum_{i=1}^{m} v_i (I_{ij})^l_\alpha \leq 0; \qquad j = 1, \ldots., n, \quad j \neq 0,$$
$$u_r, v_i \geq 0; \quad r = 1,2,\ldots,s \quad \text{and} \quad i = 1,2,\ldots m.$$

It may be noted that for every α, $E^l_\alpha \leq E^u_\alpha$ and if $\alpha_1 \leq \alpha_2$, then

$$\left[E^l_{\alpha_2}, E^u_{\alpha_2}\right] \subseteq \left[E^l_{\alpha_1}, E^u_{\alpha_1}\right].$$

2.1 Intuitionistic Fuzzy Entropy Measure

Let us consider that Atanassov's intuitionistic fuzzy set *(IFS)*, $\widetilde{A}$, over a finite non empty fixed set $X = \{x_1, x_2, \ldots, x_n\}$. The concept of the intuitionistic fuzzy entropy measure for IFSs has been characterized and discussed in [10], [11] and a set of following four properties, which an intuitionistic fuzzy entropy should satisfy, was introduced:

- **(IFS1) :** $H(\widetilde{A}) = 0$ iff $\widetilde{A}$ is a crisp set, i.e. $\mu_{\widetilde{A}}(x_i) = 0$ and $\gamma_{\widetilde{A}}(x_i) = 1$ or $\mu_{\widetilde{A}}(x_i) = 1$ and $\gamma_{\widetilde{A}}(x_i) = 0$ for all $x_i \in X$.
- **(IFS2) :** $H(\widetilde{A}) = 1$ iff $\mu_{\widetilde{A}}(x_i) = \gamma_{\widetilde{A}}(x_i)$ for all $x_i \in X$.
- **(IFS3) :** $H(\widetilde{A}) \leq H(\widetilde{B})$ if $\widetilde{A}$ is less fuzzy then $\widetilde{B}$, i.e . $\mu_{\widetilde{A}}(x_i) \leq \mu_{\widetilde{B}}(x_i)$ and $\gamma_{\widetilde{A}}(x_i) \geq \gamma_{\widetilde{B}}(x_i)$ for $\mu_{\widetilde{B}}(x_i) \leq \gamma_{\widetilde{B}}(x_i)$ or $\mu_{\widetilde{A}}(x_i) \geq \mu_{\widetilde{B}}(x_i)$ and $\gamma_{\widetilde{A}}(x_i) \leq \gamma_{\widetilde{B}}(x_i)$ for $\mu_{\widetilde{B}}(x_i) \geq \gamma_{\widetilde{B}}(x_i)$ for all $x_i \in X$.
- **(IFS4) :** $H(\widetilde{A}) = H(\overline{\widetilde{A}})$, where $\overline{\widetilde{A}}$ is complement of $\widetilde{A}$.

It may be noted that the above four axiomatic requirements, i.e, sharpness, maximality, resolution and symmetry of intuitionistic fuzzy entropy are widely accepted and have become a criterion for defining any new intuitionistic fuzzy entropy.

Corresponding to IFS $\widetilde{A}$ with n elements (intuitionistic fuzzy values) $a_i = (\mu_i, \gamma_i)$, $i = 1,2,\ldots,n$, as in [10], [11], we have the following entropy measure of IFS $\widetilde{A}$:

$$H_{sk} = \frac{1}{n}\sum_{i=1}^{n}\frac{\max count(a_i \wedge \overline{\overline{a}}_i)}{\max count(a_i \vee \overline{\overline{a}}_i)}$$

where $\max count(\widetilde{A}) = \sum_{i=1}^{n}(\mu_{\widetilde{A}}(x_i) + \pi_{\widetilde{A}}(x_i))$, $\widetilde{A} \in F(X)$. Here $F(X)$ is set of all the IFSs on X.

Also, corresponding to IFS due to De Luca $-$Termini entropy in [5], we have the following measure of IFS $\widetilde{A}$ of n elements (intuitionistic fuzzy values) $a_i = (\mu_i, \gamma_i)$, $i = 1,2,\ldots\ldots,n$:

$$E_{LT}(\widetilde{A}) = -\frac{1}{n\ln 2}\sum_{i=1}^{n}\left[\mu_i \ln\left(\frac{\mu_i}{\mu_i+\gamma_i}\right) + \gamma_i \ln\left(\frac{\gamma_i}{\mu_i+\gamma_i}\right) - \pi_i \ln 2\right].$$

The concept of De Luca $-$Termini entropy for IFSs has been properly derived in [12] from Intuitionistic fuzzy cross-entropy of the IFSs.

2.2 Intuitionistic Fuzzy Weighted Entropy Measure

The concept of Intuitionistic Fuzzy Weighting function can be seen as the decision function representing the attitude of decision maker for many real life problems such as investments, new product development, delivery personnel selections, allocation of resources and especially in multicriteria decision making and many others.

Let ϕ be a real valued function defined as

$$\phi : \varepsilon \to [0,1], \text{ where } \varepsilon = \{(\alpha,\beta) : \alpha,\beta \in [0,1], \alpha+\beta \le 1\},$$

be the set of all intuitionistic fuzzy values.

Consider two intuitionistic fuzzy values such as $\widetilde{p} = (\mu_{\widetilde{p}}, \gamma_{\widetilde{p}}), \widetilde{q} = (\mu_{\widetilde{q}}, \gamma_{\widetilde{q}}) \in \varepsilon$. ϕ is an entropy measure of IFSs, characterised as the intuitionistic fuzzy weighted entropy, if following four properties are satisfied:

- **(IFWE1) :** $\phi(\widetilde{p}) = 0$ iff $\mu_{\widetilde{p}} = 0$ and $\gamma_{\widetilde{p}} = 1$ $\big($or $\mu_{\widetilde{p}} = 1$ and $\gamma_{\widetilde{p}} = 0\big)$.
- **(IFWE2) :** $\phi(\widetilde{p}) = 1$ iff $\mu_{\widetilde{p}} = \gamma_{\widetilde{p}}$.
- **(IFWE3) :** $\phi(\widetilde{p}) \le \phi(\widetilde{q})$, if $\widetilde{p}$ is less than $\widetilde{q}$, i.e., $\mu_{\widetilde{p}} \le \mu_{\widetilde{q}}$ and $\gamma_{\widetilde{p}} \ge \gamma_{\widetilde{q}}$ for $\mu_{\widetilde{q}} \le \gamma_{\widetilde{q}}$ $\big($or $\mu_{\widetilde{p}} \ge \mu_{\widetilde{q}}$ and $\gamma_{\widetilde{p}} \le \gamma_{\widetilde{q}}$ for $\mu_{\widetilde{q}} \ge \gamma_{\widetilde{q}}\big)$.
- **(IFWE4) :** $\phi(\widetilde{p}) \le \phi(\overline{\widetilde{p}})$.

Above four axiomatic requirements, i.e., sharpness, maximality, resolution and symmetry of intuitionistic fuzzy weighted entropy are widely accepted and have become a criterion for defining any new intuitionistic fuzzy weighted entropy.

Let ϕ be a function defined as $\phi : \varepsilon \to [0,1]$, and $\widetilde{A} = \{\widetilde{a}_1, \widetilde{a}_2, ..., \widetilde{a}_n\}$ where $\widetilde{a}_i = (\mu_i, \gamma_i),\ i = 1,2, ..., n$, we have

$$\phi(\widetilde{a}_i) = \pi_i - \frac{1}{\ln 2} \cdot \sum_{i=1}^{n} \left[\mu_i \ln\left(\frac{\mu_i}{\mu_i + \gamma_i}\right) + \gamma_i \ln\left(\frac{\gamma_i}{\mu_i + \gamma_i}\right) \right]; \tag{1}$$

$\phi(\widetilde{a}_i)$ fulfils the requirement for intuitionistic fuzzy value entropy measure.

Hence, we get $E_{LT} = \frac{1}{n} \sum_{i=1}^{n} \phi(\widetilde{a}_i)$.

From above equation we can get the weighted De Luca-Termini entropy for IFSs

$$E_{WLT}(\widetilde{A}) = \sum_{i=1}^{n} w_i \phi(\widetilde{a}_i);$$

where $w_i \in (0,1]$, $i = 1,2, ..., n$ and $\sum_{i=1}^{n} w_i = 1$ i.e. $w_1 = \cdots = w_n = \frac{1}{n}$.

3 Algorithm for Sorting of Decision Making Units

As multicriteria decision making problems are defined on set of alternatives, so in this section we will discuss how to utilize the efficiency of DMUs to identify the best alternative according to some criteria. The procedure for intuitionistic fuzzy multicriteria decision making (IFMCDM) based on efficiency of DMUs and intuitionistic fuzzy weighted entropy consists of following steps:

Step 1: Take multiple inputs and multiple outputs. Estimate the efficiency of DMUs by using Fuzzy DEA model.

Step 2: Convert efficiency of DMUs to decision matrix by considering mean of efficiency interval as degree of membership of the alternatives y_j $(j = 1,2, ..., m)$ according to the criterion x_i $(i = 1,2, ..., n)$, and is denoted by intuitionistic fuzzy valued decision matrix $\widetilde{M} = [\widetilde{m}_{ij}]_{n \times m}$, where $\widetilde{m}_{ij} = (\mu_{ij}, \gamma_{ij})$. Here μ_{ij}, γ_{ij} are the degree of membership and non-membership of the alternatives.

Step 3: Make use of the principle of minimum entropy value to get the weight vector, which is defined as

$$\min E_w = \sum_{j=1}^{m} E_w\left(\widetilde{A}_j\right) = \sum_{j=1}^{m} \sum_{i=1}^{n} w_i \phi(\widetilde{m}_{ij});$$

such that

$$\begin{cases} K_w, \\ w_1 + w_2 + ... + w_n = 1, \\ w_i \geq \eta\ (i = 1,2, ..., n), \end{cases}$$

where K_w is the set of known information about the weight vector, $\widetilde{A}_j$ is the estimation given by decision maker and η is a small positive real number.

After calculating minimum value of E_w, we calculate optimal weight vector, which is given by

$$w^* = \arg\min E_w.$$

Step 4: Amassed the estimation of alternatives by intuitionistic fuzzy weighted averaging operator ($IFWA_w$) or intuitionistic fuzzy weighted geometric operator ($IFWG_w$).

Step 5: Final and most important step is to rank the alternatives y_j ($j = 1, 2, ..., m$) and select the best one in accordance with the comparison method which is given by the definition 4.

4 Illustrative Example

Let us consider an example related to a software company, searching the best supplier for one of its most important software used in assembling of Laptops.

Table 1. Data Table consisting of two fuzzy inputs and two fuzzy outputs

Decision Making Units	Supplier A	Supplier B	Supplier C	Supplier D	Supplier E
I/P-1	4, 3.5, 4.5	2.9, 2.9, 2.9	4.9, 4.4, 5.4	4.1, 3.4, 4.8	6.5, 5.9, 7.1
I/P-2	2.1, 1.9, 2.3	1.5, 1.4, 1.6	2.6, 2.2, 3.0	2.3, 2.2, 2.4	4.2, 3.6, 4.6
O/P-1	2.6, 2.4, 2.8	2.2, 2.2, 2.2	3.2, 2.7, 3.7	2.9, 2.5, 2.3	5.1, 4.4, 5.8
O/P-2	4.1, 3.8, 4.4	3.5, 3.3, 3.7	5.1, 4.3, 5.9	5.7, 5.5, 5.9	7.4, 6.5, 8.3

Efficiency of DMUs is calculated by using DEA model. For $\alpha = 0$, we have

$$\max \quad E_0^u = u_1 O_{10}^u + u_2 O_{20}^u$$

such that

$$v_1 (I_{1o})^l + v_2 (I_{2o})^l = \widetilde{1};$$
$$u_1 O_{1o}^u + u_2 O_{2o}^u - v_1 I_{1o}^l - v_2 I_{2o}^l \leq 0;$$

and

$$u_1 O_{11}^l + u_2 O_{21}^l - v_1 I_{11}^u - v_2 I_{21}^u \leq 0;$$
$$u_1 O_{12}^l + u_2 O_{22}^l - v_1 I_{12}^u - v_2 I_{22}^u \leq 0;$$
$$u_1 O_{13}^l + u_2 O_{23}^l - v_1 I_{13}^u - v_2 I_{23}^u \leq 0;$$
$$u_1 O_{14}^l + u_2 O_{24}^l - v_1 I_{14}^u - v_2 I_{24}^u \leq 0;$$
$$u_1 O_{15}^l + u_2 O_{25}^l - v_1 I_{15}^u - v_2 I_{25}^u \leq 0.$$

On substituting all values from Table 1, we get upper bound of efficiency when $\alpha = 0$. In the similar manner we get the other efficiencies of DMUs as follows in Table 2:

Table 2. Efficiency of DMUs

Decision Making Units	Supplier A	Supplier B	Supplier C	Supplier D	Supplier E
$\alpha = 0$	0.654, 1	0.836, 1	0.571, 1	0.855, 1	0.638, 1
$\alpha = 0.25$	0.702, 1	0.908, 1	0.642, 1	0.943, 1	0.735, 1
$\alpha = 0.50$	0.758, 0.963	0.99, 1	0.716, 1	1, 1	0.845, 1
$\alpha = 0.75$	0.807, 0.904	1, 1	0.791, 0.932	1, 1	0.969, 1
$\alpha = 1.00$	0.855, 0.855	1,1	0.861, 0.861	1,1	1,1

Coversion of Efficiency DMUs to Decision Matrix Table

Looking at the efficiency interval, we consider mean of efficiency interval as degree of membership of the alternatives y_j (A, B, C, D and E), satisfying the criterion x_i ($\alpha = 0, 0.25, 0.50, 0.75, 1$). The intuitionistic fuzzy index $\pi_{ij} = 1 - \mu_{ij} - \gamma_{ij}$ shows the decision maker's hesitation of the alternatives y_j with respect to criterion x_i and is zero whenever alternatives $y_j = 1$. Therefore, the decision matrix $\widetilde{M}$ obtained from efficiency of DMUs is given by

$$\widetilde{M} = \begin{pmatrix} (.827,.173) & (.918,.082) & (.785,.215) & (.927,.073) & (.819,.181) \\ (.851,.149) & (.954,.046) & (.821,.179) & (.971,.029) & (.867,.133) \\ (.860,.103) & (.995,.005) & (.858,.142) & (1,0) & (.922,.078) \\ (.855,.049) & (1,0) & (.861,.071) & (1,0) & (.984,.016) \\ (.855,0) & 1,0) & (.861,0) & (1,0) & (1,0) \end{pmatrix}$$

Let K_w the set of known information about the weight vector given by:

$$K_w = \left\{ \begin{array}{l} w_1 \leq 0.3, 0.1 \leq w_2 \leq 0.2, 0.2 \leq w_3 \leq 0.5, 0.1 \leq w_4 \leq 0.3, w_5 \leq 0.4, \\ w_3 - w_2 \geq w_5 - w_4, w_4 \geq w_1, w_3 - w_1 \leq 0.1 \end{array} \right\}$$

By using (1), we get the De Luca - Termini entropy of the intuitionistic fuzzy values as under:

$$\begin{pmatrix} 0.6645 & 0.4092 & 0.7509 & 0.3770 & 0.6882 \\ 0.6073 & 0.2691 & 0.6779 & 0.1893 & 0.5656 \\ 0.5095 & 0.0454 & 0.5894 & 0 & 0.3951 \\ 0.3708 & 0 & 0.4301 & 0 & 0.1183 \\ 0.1450 & 0 & 0.1390 & 0 & 0 \end{pmatrix}$$

Therefore,

$$E_w = \sum_{j=1}^{5} E_w\left(\widetilde{A}_j\right) = \sum_{j=1}^{5}\sum_{i=1}^{5} w_i \phi(\widetilde{m}_{ij})$$
$$= 2.8898\, w_1 + 2.3092\, w_2 + 1.5394 w_3 + 0.9192\, w_4 + 0.2840\, w_5.$$

Hence, we have the following linear programming problem:

$$\min E_w = 2.8898\, w_1 + 2.3092\, w_2 + 1.5394 w_3 + 0.9192\, w_4 + 0.2840\, w_5$$

subject to

$$\left\{ \begin{array}{l} w_1 \leq 0.3, 0.1 \leq w_2 \leq 0.2, 0.2 \leq w_3 \leq 0.5, 0.1 \leq w_4 \leq 0.3, w_5 \leq 0.4, \\ -w_2 + w_3 + w_4 - w_5 \geq 0, -w_1 + w_4 \geq 0, -w_1 + w_3 \leq 0.1, \\ w_1 + w_2 + \ldots\ldots + w_n = 1, \\ w_i \geq 0.001\, (i = 1, 2, 3, 4, 5). \end{array} \right.$$

Its optimal solution is $w_1 = 0.1$, $w_2 = 0.1$, $w_3 = 0.2$, $w_4 = 0.25$, $w_5 = 0.35$.

Now apply either $IFWA_w$ or $IFWG_w$ operator (Ref. Definition 2 and 3). Here we have applied $IFWG_w$ operator to get

$$\widetilde{a}_1 = (0.8527, 0.0671), \quad \widetilde{a}_2 = (0.9858, 0.0142), \quad \widetilde{a}_3 = (0.8484, 0.0888),$$
$$\widetilde{a}_4 = (0.9895, 0.0105), \quad \widetilde{a}_5 = (0.9469, 0.0530).$$

By applying Definition 4, we calculate score function $s(a_j)\,(j = 1,2,3,4,5)$,

$$s(a_1) = 0.7856, \quad s(a_2) = 0.9716, \quad s(a_3) = 0.7596,$$

$$s(a_4) = 0.9790, \quad s(a_5) = 0.8939.$$

Therefore, we can say that alternative D is best choice and the optimal ordering is $y_4 > y_2 > y_5 > y_1 > y_3$, i.e,

$$D > B > E > A > C.$$

5 Conclusion and Scope for Future Work

Under the new algorithm proposed in section 3, the sorting of decision making unit in data envelopment analysis has been accomplished and an optimal ranking order has been found out with the help of intuitionistic fuzzy weighted entropy according to minimum entropy model. The efficiency of the proposed methodology may be applied in regard of information measure for pattern recognition, medical diagnosis, and image segmentation.

Acknowledgements. The authors are thankful to anonymous reviewers for their valuable comments and suggestions.

References

1. Atanassov, K.: Intuitionistic fuzzy sets. Fuzzy Sets and Systems 20, 87–96 (1986)
2. Atanassov, K.: More on intuitionistic fuzzy sets. Fuzzy Sets and Systems 33, 37–46 (1989)
3. Atanassov, K., Pasi, G., Yager, R.R.: Intuitionistic fuzzy interpretations of multi-criteria multiperson and multi-measurement tool decision making. International Journal of Systems Science 36, 859–868 (2005)
4. Charnes, A., Cooper, W.W., Rhodes, E.: Measuring the Efficiency of Decision Making Units. European Journal of Operational Research 2, 429–444 (1978)
5. de Luca, A., Termini, S.: A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory. Information Control 20, 301–312 (1972)
6. Li, D.F.: Multiattribute decision making models and methods using intuitionistic fuzzy sets. Journal of Computers and System Sciences 70, 73–85 (2005)
7. Liu, H.W., Wang, G.J.: Multi-criteria decision making methods based on intuitionistic fuzzy sets. Europian Journal of Operational Research 179, 220–233 (2007)
8. Noura, A.A., Saljooghi, F.H.: Ranking Decision Making units in Fuzzy DEA using Entropy. Applied Mathematical Sciences 3, 287–295 (2009)
9. Sengupta, J.K.: A fuzzy system approach in data envelopment analysis. Computers and Mathematics with Applications 49, 259–266 (2005)
10. Szmidt, E., Kacprzyk, J.: Entropy for intuitionistic fuzzy sets. Fuzzy Sets and Systems 118, 467–477 (2001)

11. Szmidt, E., Kacprzyk, J.: Using intuitionistic fuzzy sets in group decision making. Control and Cybernetics 31, 1037–1053 (2002)
12. Vlachos, I.K., Sergiadis, G.D.: Intuitionistic fuzzy information applications to pattern recognition. Pattern Recognition Letters 28, 197–206 (2007)
13. Wu, J.-Z., Zhang, Q.: Multicriteria decision making method based on intuitionistic fuzzy weighted entropy. Experts Systems with Applications 38, 916–922 (2011)
14. Xu, Z.S., Yager, R.R.: Dynamic intuitionistic fuzzy multi-attribute decision making. International Journal of Approximate Reasoning 48, 246–262 (2008)

Reliability Quantification of an OO Design -Complexity Perspective-

A. Yadav and R.A. Khan

Department of Information Technology
Babasaheb Bhimrao Ambedkar University, Lucknow, India
{amitabha.engg,khanraees}@yahoo.com

Abstract. Object oriented design and development are popular conceptions in today's software development scenario. Object oriented design supports design principals such as inheritance, coupling, cohesion and encapsulation. The proposed research work will deliver a mechanism for reliability estimation of object oriented design in respect of complexity perspective. The four OO design metrics namely Inheritance metric complexity perspective (IM_C), coupling metric complexity perspective (CM_C), cohesion metric complexity perspective (C_OM_C) and encapsulation metric complexity perspective (EM_C) are proposed for each of object oriented design constructs such as inheritance, coupling, cohesion and encapsulation respectively. The paper also proposed complexity and reliability estimation models. On the basis of proposed metrics a multiple regression equation has been established for computing the complexity of design hierarchies. Complexity is inversely affects reliability of object oriented designs. Again a multiple regression equation has been established to compute reliability in respect of complexity. Comparative analysis among metric and model values has been done in this paper.

Keywords: Complexity, Reliability, Estimation, Design, Quantification.

1 Introduction

Software spread over our modern society very rapidly over last two decades. Nothing can be operated without the use of software like cameras, VCRs, televisions, car engines, vending machine etc. Software is a systematic representation and processing of human knowledge [1]. Today, every thing is depended on computer software. The dependency on software increases failures, when software does not execute reliably. Software executes reliably when it does what it is supposed to do. Software reliability is still a challenging problem that affects software producers from small developer teams to big vendors [2]. The increasing complexity of software systems makes reliability more difficult to handle. In addition, most software developers have insufficient knowledge of reliability aspect and use immature reliability assurance techniques for preventing reliability defects across the entire software development life cycle. This worsens the problem of reliable software engineering [1].

There are several approaches to makes the system highly reliable. Among several approaches object oriented design is one of the important approach to estimates

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 577–585.
springerlink.com

reliability in complexity perspective. Object oriented design herald itself as an important tool for solving most of the software problems [3]. In an object oriented approach, the data is treated as the most important element and it cannot flow freely around the system. Restrictions are placed that can manipulate the data [4]. Increase in the size of code, increases unnecessary effort and complexity. Complexity of the software increases with error handling functions. High complexity of software usually produces software with sever faults [5]. High complexity decreases reliability of software. However, software faults vary considerably with respect to their severity. A failure caused by a fault may lead to a whole system crash or an inability to open a file [6]. Therefore, there is a need to develop an approach that can be used to identify classes that are prone to have serious faults.

From the foregoing discussion, it seem that minimizing unwanted complexity early in the development life cycle leads to the development of high reliability end products. A metric based approach may be used to evaluate complexities and their extension for object oriented design. On the bases of the approach, reliability of the object oriented software at early phase of development can be improved. None of such an approach is available to be used in early stage of development life cycle and there is in high demand to develop a metric based complexity perspective approach for object oriented software to be used in design phase. The object oriented paradigms have a capability to adjust complexity of object oriented design by maintaining reliability of software [7]. The proposed research work will deliver a mechanism for reliability estimation of object oriented design in respect of complexity perspective. Complexity of object oriented design can be maintained by adjusting object oriented constructs such as inheritance, coupling, cohesion and encapsulation. Highly complex system makes the system unreliable. Reliability of a software system increases by making the system less complex.

Section 2 presents four proposed metrics. Complexity estimation model and reliability estimation model is covered in section 3 and 4 respectively. Comparative analysis is done in section 5. Findings are included in section 6. At last paper is concluded in section 7.

2 Proposed Metrics

Many metrics were proposed by different researchers and practitioners for object oriented software. Metrics are the quantitative measure of the degree to which a system component or process possess a given attribute [8]. The main aim of software metrics is the recognition and measurement of the necessity parameters that affects software development. Software metrics are broadly classified into product metrics or process metrics [9]. Software metrics are the level of quantifiable measurement for which a system element or procedures possess a given software dimension. Software reliability metrics are used for software reliability rating and confidence and they are very important for software reliability because of quantification, cost, schedule, reliability, forecasting and validation [10-11]. Software reliability metrics are useful to know the probability of software failure or the rate at which software errors will occurs [12]. The following metrics are developed during the research:

- Inheritance metric complexity perspective (IM_C) [13]
- Coupling metric complexity perspective (CM_C) [14]
- Cohesion metric complexity perspective (C_OM_C) [15]
- Encapsulation metric complexity perspective (EM_C) [16]

3 Complexity Estimation Model

In order to establish a relationship between design constructs and reliability attribute complexity, the respective influence of design constructs on complexity is being examined on the basis of critical literature survey [17]. It was observed that each of the design constructs affects complexity and complexity affects reliability of object oriented software. The extensive review of object oriented development literature reveals that object oriented constructs negatively or positively affects software complexity and complexity negatively affects software reliability. Researcher uses object oriented design constructs to estimate the complexity OO software. Object oriented design constructs such as high inheritance and coupling positively affects complexity of software [18-20]. In same way high cohesion and encapsulation decreases complexity of software. The proposed metrics are being used for estimating complexity of object oriented design using complexity estimation model (CEM). A multiple linear regression has been established to get coefficients. The multiple linear regressions establish a relationship between dependent variables and multiple independent variables. Thus, the multiple regression equation takes the form as follows:

$$Y= \alpha_0+\beta_1X_1+ \beta_2X_2..... \beta_nX_n \quad (1)$$

Where Y is the dependent variable α and β are the regression coefficients, and X_s are independent variables. Putting equation 1, 2, 3 and 4in equation 5 the generated equation will be as follows:

$$C= \alpha+\beta\ (IM_C) +\lambda\ (CM_C) +\gamma\ (C_OM_C) +\delta\ (EMc) \quad (2)$$

Where, C is complexity which is dependent variable, IM_C (Inheritance metric complexity perspective), CMc (Coupling metric complexity perspective), C_OM_C (cohesion metric complexity perspective), EM_C (Encapsulation metric complexity perspective) are the metrics which are worked as independent variables and α, β, λ, δ ,γ are treated as regression coefficients. Complexity depends on design constructs such as inheritance, coupling, cohesion and encapsulation. Five different equations are set for five different case studies. On the basis of these case studies data regression coefficients for object oriented design constructs such as inheritance, coupling, cohesion and encapsulation are computed as α=0.8012, β=-0.8079, γ= -0.4374, δ=-0.6221,λ=0.1421 respectively. Putting values of α, β, γ, δ, λ in equation 2, following equation will be generated

$$C= 0.8012- 0.8079*(IM_C) +0.4374*\ (CM_C) -0.6221*\ (C_OM_C) +0.1421*\ (EM_C) \quad (3)$$

4 Reliability Estimation Model

In order to establish a relationship between reliability and complexity, the respective influence of relationship between complexity and reliability are being examined on the basis of literature survey [21]. It was observed that complexity and reliability are closely related with each other and complexity negatively affects reliability of object oriented software [22]. Researcher uses the complexity of object oriented design to estimate reliability of software. Highly complex software decreases the reliability of object oriented design [23]. The developed equation 1 is being used for estimating reliability of object oriented design. A multiple linear regression has been established to get coefficients. The multiple linear regressions establish a relationship between dependent variables and multiple independent variables [24]. Thus, the multiple regression equation takes the form as follows:

$$Z= \lambda_0+\Gamma_1 X_1+ \Gamma_2 X_2 \ldots\ldots\ldots\ldots\ldots\ldots \Gamma_n X_n \quad (4)$$

Where Z is the dependent variable λ and Γ are the regression coefficients, and X_s are independent variables.

$$R= \alpha+\beta\ (C) \quad (5)$$

Where R is reliability which is dependent variable, C is complexity which worked as independent variables and α, β is treated as regression coefficients. Reliability depends on complexity. Five different equations are set for five different case studies. On the basis of these case studies data, regression coefficients for complexity are computed as α=1.041, β= -0.263. Putting values of α, β, in equation 5, following equation will be generated

$$R= 1.041-0.263\ (C) \quad (6)$$

5 Comparative Analysis

Comparison between proposed metrics, reliability and complexity is shown in following tables and figures:

- Analysis of OO design metrics, complexity and reliability values for the designs before complexity minimization.

Table 1. Metric, complexity and reliability values for the designs

Metric / Designs	IM_C	CM_C	C_OM_C	EM_C	Complexity	Reliability
Design 1	0.381	0.968	0.351	0.967	0.8358	0.8212
Design 2	0.289	0.821	0.313	0.923	0.8633	0.8140
Design 3	0.321	0.953	0.334	0.897	0.8784	0.8100
Design 4	0.246	0.839	0.284	0.931	0.9251	0.7977
Design 5	0.189	0.973	0.471	0.835	0.8997	0.8044

- Analysis of metrics, complexity and reliability values after complexity minimization

Table 2. Metrics, complexity and reliability comparison after Complexity minimization

Metric / Designs	IM_C	CM_C	C_OM_C	EM_C	Complexity	Reliability
Design 1	0.272	0.732	0.854	0.989	0.5109	0.9066
Design 2	0.279	0.721	0.892	0.931	0.4685	0.9178
Design 3	0.196	0.569	0.859	0.996	0.4847	0.9135
Design 4	0.216	0.782	0.759	0.957	0.6326	0.8746
Design 5	0.135	0.729	0.853	0.887	0.6064	0.8815

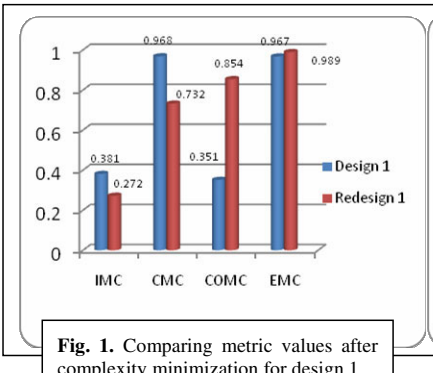


Fig. 1. Comparing metric values after complexity minimization for design 1

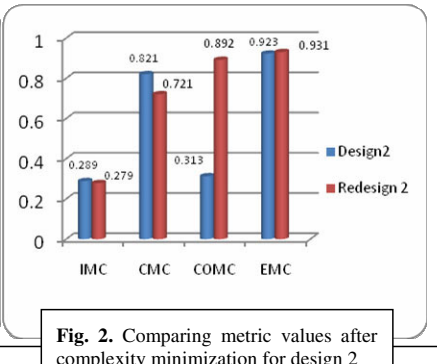


Fig. 2. Comparing metric values after complexity minimization for design 2

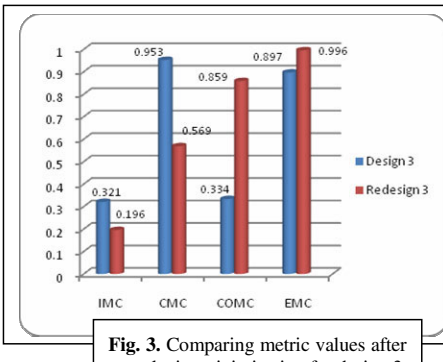


Fig. 3. Comparing metric values after complexity minimization for design 3

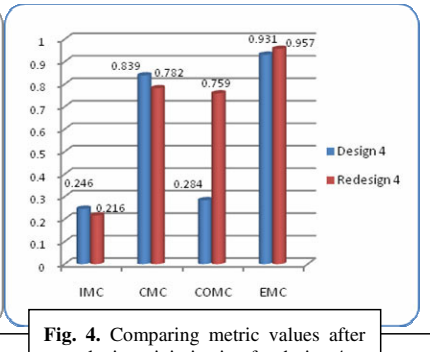


Fig. 4. Comparing metric values after complexity minimization for design 4

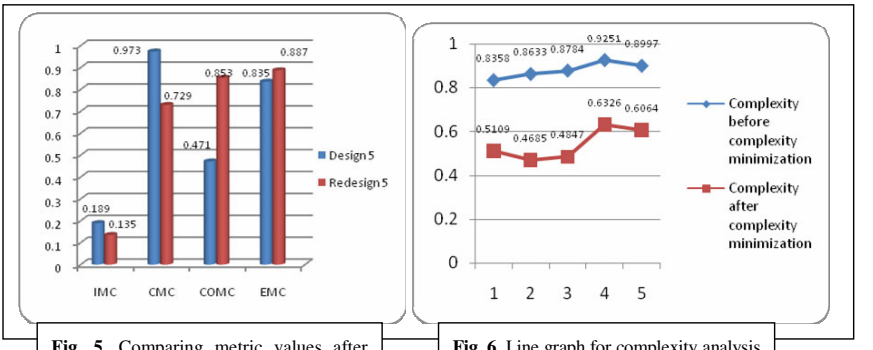


Fig. 5. Comparing metric values after complexity minimization for design 5

Fig. 6. Line graph for complexity analysis after before complexity minimization

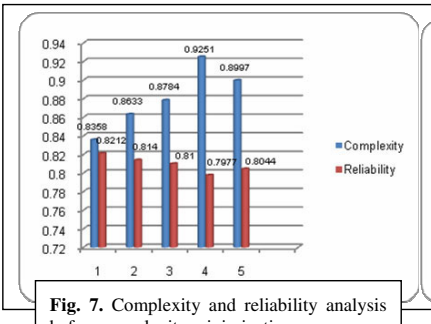


Fig. 7. Complexity and reliability analysis before complexity minimization

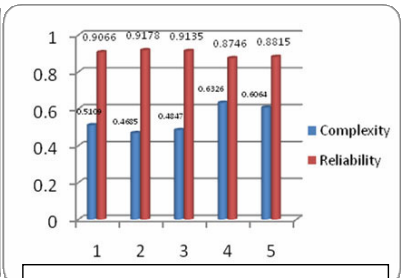


Fig. 8. Analysis of complexity and reliability after complexity minimization

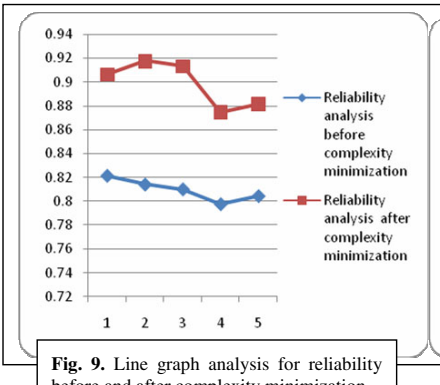


Fig. 9. Line graph analysis for reliability before and after complexity minimization

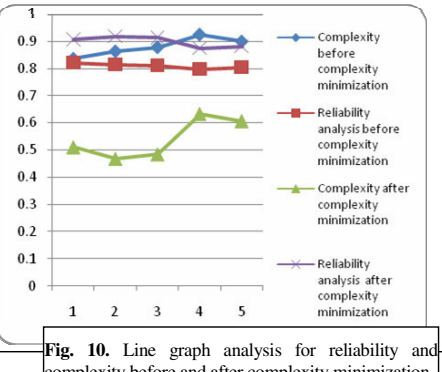


Fig. 10. Line graph analysis for reliability and complexity before and after complexity minimization

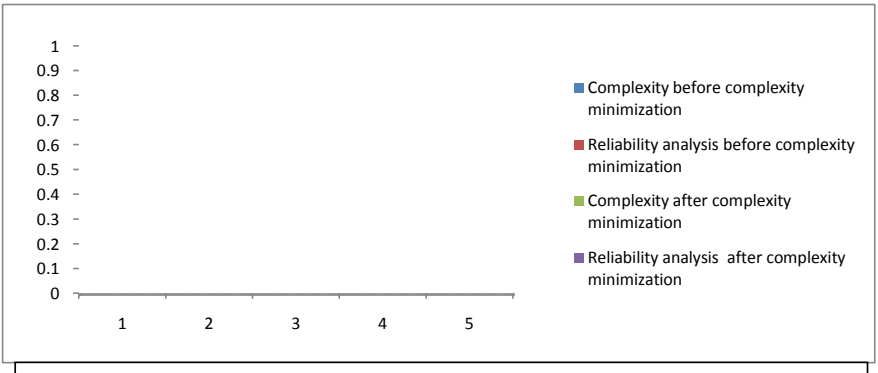


Fig. 11. Analysis for reliability and complexity before and after complexity minimization

6 Findings

Complexity is key factor to reliability. High complex software decreases the reliability of software. Some of the major findings are as given below:

- The four metrics Inheritance metric complexity perspective (IM_C), coupling metric complexity perspective (CM_C), cohesion metric complexity perspective (C_OM_C) and encapsulation metric complexity perspective (EM_C) are proposed for each of object oriented design constructs such as inheritance, coupling, cohesion and encapsulation respectively.
- Complexity and reliability estimation models have been proposed.
- Comparative analysis between metric values, reliability and complexity is the major part of the proposed work that has been done shown in above tables and figures.
- It is found after analysis of older values of metrics that, metrics IM_C and CM_C has more values than newer once. In the same way older values of C_OM_C and EM_C have fewer values than newer once. Before, complexity minimization, complexity of the design has higher values than after complexity minimization. Same has for reliability of the design has fewer values than after minimization. Hence, it is concluded that higher the complexity, lower is the reliability of object oriented design.
- It is found after analysis of metrics, complexity and reliability values after complexity minimization is that complexity has lesser values than reliability. Hence, lesser the complexity, higher the reliability of object oriented design.
- Metrics values are compared with each other before and after complexity minimization for all five designs. It is identified that values of complexity before complexity minimization has greater values for IM_C and CM_C and lesser values for C_OM_C and EM_C than after complexity minimization for the designs. Hence, complexity values are minimized for five designs shown in tables and figures.
- Complexity has been analyzed for five designs after and before complexity minimization for the designs shown in table 1 and table 2 and figures 6 and 11.

- It is found that before complexity minimization that reliability is lesser than complexity values. Hence, higher the complexity, lesser is the reliability of designs. Table 1 and figure 7 depict the comparative analysis between them.
- It is found that after complexity minimization, complexity values are lesser than the values of reliability. Hence, lesser the complexity greater is the reliability of the designs. Table 2 and figure 8 depict the analysis after complexity minimization for the designs.
- Reliability analysis after complexity minimization is shown in table 2 and represented in figure 9 and 10. It is found that reliability values are high after complexity minimization.
- Reliability and complexity analysis before & after complexity minimization for the designs are given in table 1 and table 2 and figure 11. It is found that complexity decreases and reliability of designs increases.

7 Conclusion

The four OO design metrics namely Inheritance metric complexity perspective (IM_C), coupling metric complexity perspective (CM_C), cohesion metric complexity perspective (C_OM_C) and encapsulation metric complexity perspective (EM_C) are proposed for each of object oriented design constructs such as inheritance, coupling, cohesion and encapsulation respectively. Complexity and reliability estimation models have also proposed for quantifying complexity and reliability of the design respectively. Comparative analysis between metric values, reliability and complexity is the major part of the proposed work has been done. It is found after analysis that high complexity in the design decreases the reliability of the design. After complexity minimization it is found that values of complexity is less than the values of reliability. Hence, complexity of object oriented designs can be controlled by controlling design constructs. Hence tables and graphical representation depicts the comparative analysis between complexity and reliability after and before complexity minimization.

References

1. Michael, R.L.: Software Reliability Engineering: A Roadmap. Future of Software Engineering, 153–170 (2007) ISBN: 0-7695-2829-5
2. Zainab, A.-R., Mohammad, R., Alaa, F.S., Sulieman, B.A., Saleh, A.O.: A New Software Reliability Growth Model: Genetic-Programming-Based Approach. Int. J. Software Engineering and Applications 4, 476–481 (2011)
3. Arora, D., Khanna, P., Tripathi, A., Sharma, S., Shukla, S.: Software Quality Estimation through Object Oriented Design Metrics. Int. J. Computer Science and Network Security 11(4), 100–104 (2011)
4. Sharygina, N., Browne, C.J., Kurshan, P.R.: A Formal Object-Oriented Analysis for Software Reliability: Design for Verification, pp. 1–15 (2011)
5. Philippe, W., Lionel, J.: Complex System Reliability Modeling with Dynamic Object Oriented Bayesian Networks (DOOBN). Reliability Engineering and System Safety 91, 149–162 (2006)

6. Tsantalis, N., Chatzigeorgiou, A.: Predicting the Probability of Change in Object-Oriented Systems. IEEE Transactions on Software Engineering 31, 601–614 (2005)
7. Cristescu, M., Ciovica, L.: Estimation of the Reliability of Distributed Applications. Informatica Economică 14, 19–29 (2010)
8. Gaudan, S., Motet, G., Auriol, G.: A New Structural Complexity Metrics Applied to Object Oriented Design Reliability Assessment, http://www.lesia.insatoulouse.fr/~motet/papers/2007_ISSRE_GMA.pdf
9. Mills, E.E.: Software Metrics. SEI Curriculum Module SEI-CM-12-1.1. Software Engineering Institute, 1–43 (1988)
10. Li, H., Lu, M., Li, Q.: Software Reliability Metrics Selecting Method Based on Analytic Hierarchy Process. In: Sixth International Conference on Quality Software, QSIC 2006, October 27-28, pp. 337–346 (2006) ISSN: 1550-6002, ISBN: 0-7695-2718-3
11. Offutt, J., Alexander, R.: A fault Model for Subtype Inheritance and Polymorphism. In: Symposium, Software Reliability Engineering, pp. 84–93 (2001)
12. Li., F., Yi, T.: Apply Page Rank Algorithm to Measuring Relationship's Complexity. IEEE, 914–917 (2008) ISBN: 9780769534909
13. Yadav, A., Khan, R.A.: Measuring Design Complexity: An Inherited Method Perspective. ACM SIGSOFT Software Engineering Notes 34, 1–5 (2009) ISSN: 0163-5948, doi:0.1145/1543405.1543427
14. Yadav, A., Khan, R.A.: Coupling Complexity Normalization Metric-An Object Oriented Perspective. In. J. of Information Technology & Knowledge Management, Impact Factor 0.475 4, 501–509 (2011)
15. Yadav, A., Khan, R.A.: Class Cohesion Complexity Metric (C3M). In: IEEE In. Conference on Computer & Communication Technology (ICCCT 2011), pp. 363–366. IEEE Explorer (2011) ISBN: 978-1-4577-1385-9
16. Yadav, A., Khan, R.A.: Development of Encapsulated Class Complexity Metric. In: In. Conference on Computer, Communication, Control and Information Technology, C3IT 2012, Academy of Technology (2012)
17. Chhillar, U., Bhasin, S.: A New Weighted Composite Complexity Measure for Object-Oriented Systems. In. J. of Information and Communication Technology Research 1, 101–108 (2011) ISSN-2223-4985
18. Zhu, Y.C.Q.: Improved Metrics for Encapsulation Based on Information Hiding. In: Conference for Young Computer Scientists, pp. 742–724. IEEE computer society (2008)
19. Scharil, N., Black, A.P., Ducasse, S.: Object oriented Encapsulation for Dynamically Typed Languages. In: OOPSLA, pp. 130–139 (2004)
20. Yadav, A., Khan, R.A.: Complexity: A Reliability Factor. In: IEEE International Advance Computing Conference (IACC 2009), Thapar, pp. 2375–2375 (2009)
21. Dallal, J.A.: Mathematical Validation of Object-Oriented Class Cohesion Metrics. In. J. of Computers 4, 45–52 (2010)
22. Yadav, A., Khan, R.A.: Reliability Estimation of Object Oriented Design. IUP Journal of System Management IX, 28–41 (2011) ISSN: 0972-6896
23. Fiondella, L., Gokhale, S.S.: Software Reliability Model with Bathtub-Shaped Fault Detection Rate. In: Reliability and Maintainability Symposium (RAMS), pp. 1–6 (2011) ISBN: 978-1-4244-8857-5
24. Mohan, K.K., Verma, A.K., Srividya, A.: Software Reliability Estimation through Black Box and White Box Testing at Prototype Level. In: Conference on Reliability, Safety and Hazard (ICRESH), pp. 517–522 (2010) ISBN: 978-1-4244-8344-0

A New Hybrid Algorithm for Video Segmentation

K. Mahesh[1] and K. Kuppusamy[2]

[1] Associate Professor, Department of Computer Sci. and Engg.
Alagappa University, Karaikudi, Tamilnadu, India
mahesh.alagappa@gmail.com
[2] Associate Professor, Department of Computer Sci. and Engg.
Alagappa University, Karaikudi, Tamilnadu, India
kkdiksamy@yahoo.com

Abstract. Video segmentation became popular and most important in the digital storage media. In this video segmentation technique, initially the similar shots are segmented, subsequently the track frames in every shots are assorted using the extracted objects of every frame which highly reduces the processing time. Effective video segmentation is a challenging problem in digital storage media. In this hybrid video segmentation technique, it yields the effective video segmentation results by performing intersection on the segmented results provided by both the frame difference method as well as consecutive frame intersection method. The frame difference method considers the key frame as background and it segments the dynamic objects whereas the frame difference method segments the static and dynamic objects by intersection of objects in consecutive frames. The new hybrid technique is evaluated by varying video sequences and the efficiency is analyzed by calculating the statistical measures and kappa coefficient.

Keywords: Video segmentation, Discrete cosine transform, k-means clustering, frame difference algorithm, Euclidean distance.

1 Introduction

Due to the rapid increase in the number of digital video documents produced, the real challenge for computer vision is the development of robust tools for their utilization. It entails tasks such as video indexing, browsing or searching. Video Segmentation is one of the most important processes performed to achieve these tasks [14]. In several video processing applications, video segmentation is performed as an important step and automatic video indexing is an essential feature in the design of a video database [6]. The primary aspect of video indexing is it has the potential to segment the video into consequential or meaningful segments [1]. Video applications such as surveillance, communications and entertainment are very crucial for industries. Among these applications, segmentation of moving object is one of the basic operations for several automated video analysis tasks [15]. One of the main challenges in computer vision is automatic comprehension of complex dynamic content of videos, such as detection, localization, and segmentation of objects and people, and understanding their interactions [9]. Image and

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 587–595.
springerlink.com

video segmentation is very beneficial in several applications for finding the regions of interest in a panorama or annotating the data [11]. MPEG-4 is a promising standard for multimedia communications. MPEG-4 provides standardized ways to encode the video and audio objects, and the scene description, which represents how the objects are structured in a scene [4].

Recent development of range-camera technology has the potential to capture the range video in an applicable frame rate and frame resolution [7]. The video segmentation is an imperative technique used for the improvement of video quality on the basis of segmentation [8]. The function of video segmentation is to segment the moving objects in video sequences [19]. Video segmentation is entirely different from single image segmentation [5]. The bad quality segments such as very blurred or shaking clips should be eliminated or recovered, because these clips often irritate the viewers [13]. Video object segmentation is an important issue in video analysis, and it has several applications namely post-production, special effects, object detection, object tracking, and video compression [3].

In video segmentation, the video is segmented into spatial, temporal, or spatiotemporal regions that are consistent in some feature space [4]. Video segmentation is an important process in image sequence analysis and its results are broadly employed for describing the motion features of scene objects, and also for coding purposes to minimize the storage requirements [17]. Different methods and algorithms have been introduced for video segmentation, where each having its own features and applications [18]. These video segmentation algorithms are classified into three categories: edge information based video segmentation, image segmentation based video segmentation and change detection based video segmentation [12].

Usually, video object segmentation is done in an interactive or supervised manner. Interactive techniques necessitate a user to define object boundaries in some key frames, which are then propagated to other frames while a user stands by to adjust errors [20]. Since unsupervised video segmentation involves a huge number of data as well as image segments suffer from noisy variations in color, texture and motion with time, segmentation is a challenging problem [10].Video segmentation is an imperative component of numerous video analysis and coding problems, such as 1) video summarization, indexing, and retrieval, 2) advanced video coding, 3) video authoring and editing, 4) enhanced motion (optical flow) estimation, 5) 3D motion and structure assessment with several moving objects [16].

2 Related Work

Yasira Beevi P and S. Natarajan [1] have proposed a video segmentation algorithm for MPEG-4 camera system by means of change detection, background registration methods and real time adaptive threshold techniques.

Panagiotis Sidiropoulos *et al.*[2] have proposed a technique, where the low-level and high-level features extracted from the visual and the aural channel have been used jointly.

Kuo Liang chungi *et al.*[3] have developed a promising predictive watershed-based video segmentation algorithm using motion vectors.

Engin Mendi, *et al.*[4] have presented a medical video segmentation and retrieval research initiative.

Onur Kucüktunc [5] has proposed a fuzzy color histogram-based shot-boundary detection algorithm devoted for content based copy detection (CBCD) applications.

3 Dynamic and Static Foreground Segmentation Using Hybrid Technique

The proposed technique segments both the dynamic and static foreground objects without considering the global motion constraints. The motion segmentation process is carried out by both the frame difference algorithm and intersection method subsequently the most common and accurate segmented objects are retrieved from both the segmented results whereas the static foreground are segmented using the intersection of consecutive frames.

3.1 Shot Segmentation

The term video commonly refers to several storage formats for storing moving pictures. The video consists of consequent sequence of frames which is a single picture or still shot run in succession to produce what appears to be seamless piece of film or video tape. Let

$$V_{M\times N} = \left\{ f_{(i)}(x,y) \mid j = 1,2,.....L\,;\, x = 0,1,2,....M-1\,;\, y = 0,1,2,....N-1 \right\}$$

be the video to be segmented where $'L'$ is the total no of frames present in the video. A shot is defined as a sequence of frames which are captured from a single camera operation. Prior to video segmentation, the shot segmentation is necessary for grouping the similar shots. In this shot segmentation similar shots are grouped together for improving the performance of the segmentation. To accomplish this task initially all the frames are partitioned into $m \times n$ patches and every patch is converted to its equivalent frequency coefficients by means of Discrete Cosine Transform (DCT) developed by Ahmed, Natarajan and Rao (1974) (i.e) DCT is applied to every patch in the frames as follows:

$$D_i(a,b) = \left(2(m\times n)^{1/2}\right)\eta_a\eta_b \sum_{p=0}^{m-1}\sum_{q=0}^{n-1} f_{i,k}(p,q)\cos\left(\frac{\pi.a}{2\mathrm{m}}(2\mathrm{p}+1)\right)\cos\left(\frac{\pi.b}{2\mathrm{n}}(2\mathrm{q}+1)\right) \tag{2}$$

$1 \le k \le P_{(i)}$ where P_i is the number of patches present in the i[th] frame. Thus all the patches are transformed and *the Euclidean distance of every patch*es of consequent frames and their total mean are calculated as follow.

$$D_j = \frac{\sum_{i=1;k=1}^{L;P\,i} \sqrt{(T_{i,k} - T_{i+1,K})}}{P(i)} \tag{5}$$

Where $1 < i, j \le L$ and $1 \le k \le P_{(i)}$. The frames belongs to the similar shots are identified based on the mean distance. The fig. 1 illustrates the process of shot segmentation. Euclidean distance of every patch in the consequent frames is calculated as sample shown in fig.1.

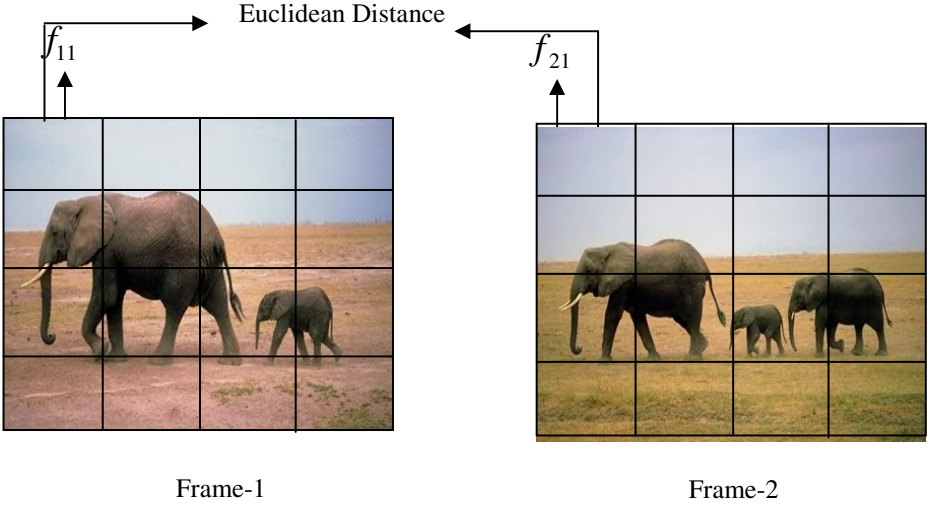


Fig. 1. Process of shot segmentation

3.2 Object Extraction

Initially the objects in every frame are identified for segmentation. Let $\delta = \{\delta_a \mid 1 < a \le A\}$ be the result of shot segmentation where 'A' is total no of shots and $\delta_a = \{ f_{aj} \mid 1 < a \le A ; 1 < j \le |\delta_a| \}$ be the set of similar shots where '$|\delta_a|$' are the total no of frames in a[th] shot in the segmented results. The initial frames

in every shot are taken as key frame for object extraction for example the f_{11} is key frame for shot δ_1 which is known as $f_{key(1)}$. Like wise each shot having its own key frames. Initially all the frames which are in RGB color format are converted to grey scale format. A RGB color is another format for color images and it represents an image with three matrices of sizes matching the image format while each matrix corresponds to one of the colors red, green and blue. [26] When we convert it into a grey scale (or "intensity") image, it depends on the sensitivity response curve of detector to light as a function of wavelength [27].

The objects are identified by the clustering process which is carried out using fuzzy k-means clustering. Finally the clustering process yields the no of objects in a frame. The overlapping objects are also identified using fuzzy k-means clustering.

3.3 Track Frame Assortment

After performing shot segmentation, the track frames of the every shot are identified using the objects of their key frame. The objects that appear simultaneously in at least two consecutive frames can be compared directly in terms of their motion so the assortment of the track frames is a required preprocessing step for segmentation this track frame selection process reduces the computational time of segmentation. The objects of the key frame are compared with the other frames of the shot for their presence in the frame.

3.4 Dynamic Foreground Segmentation Based on Frame Difference Algorithm

In the background subtraction method the key frame of every shot is consider as background. At each $\hat{f}_{aj}$ frame, the $\hat{f}_{aj}(p,q)$ pixel's value can be classified as foreground pixel if the following inequality

$$\hat{f}_{aj}(p,q) - \hat{f}_{a1}(p,q) > \lambda \tag{7}$$

Holds; otherwise $\hat{f}_{aj}(p,q)$ will be classified as background pixel value. Where $\hat{f}_{aj}(p,q)$ is the current frame pixel value, $\hat{f}_{a1}(p,q)$ is the key frame value and ' λ ' is the threshold pixel value in foreground.

3.5 Static and Dynamic Foreground Segmentation Using Intersection of Frames

The motion analysis and segmentation of dynamic objects is performed by intersection process of track frames. Initially the frames in every shot are converted to binary form.

3.5.1 Binarization

Binarization is a technique by which the gray scale images are converted to binary images. Binarization separates the foreground (text) and background information.

$$f_{\text{aj}} = \begin{cases} 1; \text{ if } & f_{\text{aj}} > \Gamma \\ 0; \text{ otherwise} & \end{cases} \tag{8}$$

Where ' Γ' is a global threshold value for binarization. After performing the binarization the consecutive frames are intersected to segment the dynamic and static objects. Let $\overline{f}_{12}$ $\overline{f}_{11}$ and $\overline{f}_{12}$ be the binarized form of frame1 and frame2 in shot1 respectively. The dynamic motion objects are found as follows

$$G2_{aj} = \overline{f}_{11} - f'_{aj} \tag{9}$$

Where as the static foreground are segmented as follows

$$f^{i}_{aj} = \overline{f}_{11} \cap \overline{f} \tag{10}$$

Like wise all the consecutive frames are intersected to achieve the static and dynamic object segmentation.

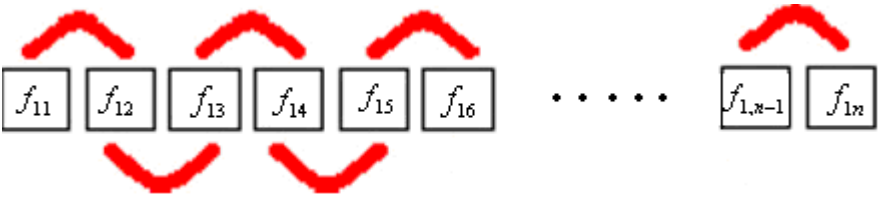


Fig. 2. Computing static and dynamic segmentation of consecutive frames

3.6 Hybridation of Segmentation Methods

Let $G1 = \{S_i \mid 1 < i \leq n\}$ and $G2 = \{S'_i \mid 1 < i \leq n\}$ be the segmented results of dynamic objects using frame difference algorithm and frame intersection method respectively. The prior said technique yields the segmented motion objects by subtracting the background and the later segmentation technique yields the motion object by intersection method.

4 Experimental Results

Our proposed video segmentation approach has been validated by experimenting with variety of video sequences. The proposed system has been implemented in Matlab (Matlab7.10).

(a)

(b)

Fig. 3. Sample similar shot segmented results

The proposed hybrid segmentation technique yields the dynamic object segmentation results by intersection of segmented results of both the frame difference algorithm and intersection methods and hence produces the better enhanced segmented results. Also the proposed system segments the static objects in every frame.

4.1 Performance Evaluation

The performance of the proposed system is evaluated by the statistical measures like sensitivity and specificity. The output of the proposed system may be positive (Segmenting the objects) or negative (non-segmenting the objects). The output of the proposed system may or may not be match with the original status of the image.

The performance is also analysis by the kappa coefficient which is as below. The table _1 represents the statistical measures of the proposed system for the different frames in a video sequence-I.

4.2 Comparative Analysis

The performance of the proposed hybrid segmentation technique is also evaluated by comparing its segmented results with that of the traditional video segmentation technique which uses background substraction method.

The table-1 and table-2 represents the comparison statistical measures of the segmentation of video-I using the proposed technique as well as the conventional method and the fig.4 illustrates the corresponding accuracy comparison graph.

Table 1. Statistical Measures of the proposed System for Video-I using the proposed system

Measures	Frame3	Frame4	Frame5	Frame7
TP	5270	5814	8914	6731
TN	92966	93355	90323	93650
FP	3140	2207	2139	995
FN	1756	2841	8285	5219
TPR or Sensitivity	75.00712	67.17504	51.82859	56.32636
FPR	3.27	2.31	2.31	1.05
Accuracy	95.25	95.16	90.49	94.17
Specificity or True Negative rate	96.73	97.69	97.69	98.95
Positive Predictive value	62.66	72.48	80.65	87.12
Negative Predictive value(NPV)	98.15	97.05	91.6	94.72
False discovery rate	37.34	27.52	19.35	12.88
Mathews Correlation Coefficient	0.66	0.67	0.6	0.67
Kappa Coefficient	0.94149	.957513	0.963626	0.980707

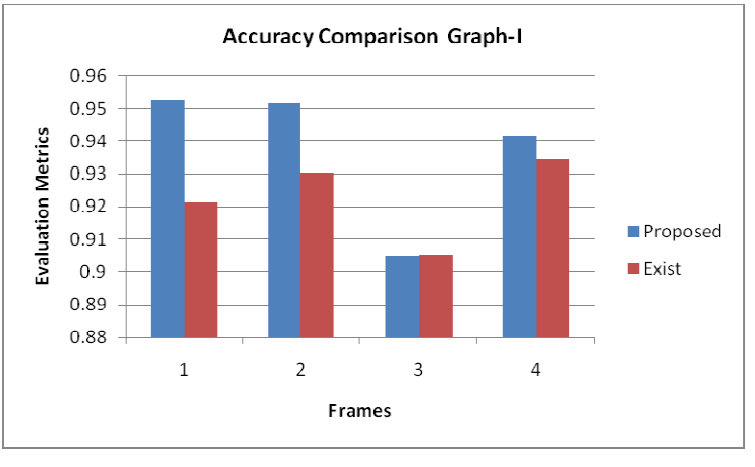


Fig. 4. Accuracy Comparison graph-I

5 Conclusion

In this paper, we have proposed a hybrid video segmentation technique to segment both the static and dynamic objects. This is has intended to overcome the existing frame difference based segmentation techniques. The segmentation technique based on frame difference algorithm segmented the objects by considering the key frame as background which only produced the motion difference from key frame with remaining frames but the proposed technique also considered the consecutive frame differences by using the consecutive frame intersection method and hence provided better result.

References

[1] A Hidden Markov Model Framework for Video Segmentation Using Audio and Image Features, http://cite-seerx.ist.psu.edu/viewdoc/download?doi=10.1.1.20...rep

[2] Sidiropoulos, P., Mezaris, V., Kompatsiaris, I., Meinedo, H., Bugalho, M., Trancoso, I.: Video scene segmentation system using audio visual features. In: Workshop on Image Analysis for Multimedia Interactive Services, WIAMIS (2010)

[3] Fathi, A., Balcan, M.F., Ren, X., Rehg, J.M.: Combining Self Training and Active Learning for Video Segmentation. In: British Machine Vision Conference (2011)

[4] Erdem, Ç.E., Sankur, B.: Performance evaluation metrics for object-based video segmentation. In: Proceeding of IEEE International Conference on Image Processing (2001)

[5] Khan, S., Shah, M.: Object Based Segmentation of Video Using Color, Motion and Spatial Information. In: Proceedings of the Conference on IEEE Computer Society, vol. 2(1) (2003)

[6] Murmu, K., Kumar, V.: Wavelet Based Video Segmentation and Indexing. EE678 Wavelets Application Assignment (April 2005)

[7] Haindl, M., Zid, P., Holub, R.: Range video segmentation. In: Proceedings of the IEEE 10th International Conference on Information Science, Signal Processing and their Applications (2010)

[8] Yadav, R.K., Sharma, S., Verma, J.S.: Deformation and Improvement of Video Segmentation Based on morphology Using SSD Technique. IJCTA 2(5), 1322–1327 (2011)

[9] Lezama, J., Alahari, K., Sivic, J., Laptev, I.: Track to the Future: Spatio-temporal Video Segmentation with Long range Motion Cues. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2011)

Using Modularity with Rough Information Systems

Ahmed T. Shawky[1], Hesham A. Hefny[1], and Ashraf H. Abd Elwhab[2]

[1] Computer Sciences and Information Department, Institute of Statistics and Research, Cairo University, Egypt
ah_taisser@yahoo.com, hehefny@hotmail.com
[2] Computer Sciences and Systems Department, Electronics Research Institute, Cairo, Egypt
awahab@ad.gov.eg

Abstract. We are looking forward to propose a novel technique, which depends on using modular techniques and integration between fuzzy set concepts and rough set theory in mining rough systems. In this research We propose a set of algorithms For a novel model allows introducing modularity mechanism; by introduce decision grouping mechanism for getting the optimizing decision. This approach provides flexibility in decision making verifies all decision standards and determines decision requirements, through modularizing rough information system, extraction of rough association rules and developing mechanisms for decision grouping.

Keywords: Rough sets, Fuzzy sets, modularity, Data mining.

1 Introduction

One of the newest approaches, which are suggested to improve the performance of decision making process based on relational databases, is the integration between fuzzy rules and rough set theory. Such an approach adopts rough set theory with applying fuzzy rules to use in data mining.

The integration of Fuzzy Set theory and Rough Set theory can achieve the flexibility of manipulation of uncertainty, and the modularity techniques overcome the problem of complexity as they split the rough decision table to smaller decision tables, which simplify reduction process by decreasing the number of attributes.

2 Rough Sets

The rough set theory, proposed by Pawlak (1982, 1996), can serve as a new mathematical tool for dealing with data classification problems. In this research we assume that data are presented in the form of decision tables. So, Rough set theory is the pest tool for dealing with data for helping decision makers. The decision table as shown later in this section consists of Rows and columns. Rows of the decision table represent cases, while columns represent variables. The set of independent variables are called conditional attributes and a dependent variable is called a decision attribute. In this section some concepts of rough set theory will be represented as decision table, indescribability relation, and decision rules.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 597–608.
springerlink.com

2.1 Information Systems

Information systems as presented by Pawlak is a table, where each row represents a case and every column represents a property of each case, the attribute may be also supplied by a human expert or user.

Table 1. An example of Decision Table

Object	Conditional Attributes			Decision
U	Age	Height	Gender	Accepted
X1	Young	Tall	Male	Yes
X2	Baby	Tall	Female	Yes
X3	Young	Tall	Female	Yes
X4	Old	Medium	Female	No
X5	Baby	Short	Male	Yes
X6	Old	Medium	Male	NO

An information system as a basic concept in rough set theory provides a convenient framework for the representation of objects in terms of their attribute values. An information system S is a pair (U, A), where U is a non-empty, finite set of objects and is called the universe and A is a non-empty, finite set of attributes. V is the set of all attribute values, such as V_a: $U \times A \rightarrow V$, $\forall x \in U$. In this example U = {X1, X2, X3, X4, X5, X6}, A= {Age, Height, Gender, Accepted}, and V(X1, Age) = Young [4].

2.2 Indiscernibility Relation

Indiscernibility relation is one of fundamentals in rough set theory, called elementary set and denoted by $[x]_B$. It may be computed by using attribute-value pair blocks as follows: For $B \subseteq A$ and $x, y \in U$, the Indiscernibility relation IND(B) is a relation on U, $(x, y) \in IND(B)$ if and only if $V_{(x, a)}=V_{(x, a)}$ $\forall a \subseteq B$. For example if t = (Gender, Male) then [t] = {X1, X5, X6} [1], [2], [6].

2.3 Reducts

For $B \subseteq A$, B is called reduct if and only if:

- $B^* = A^*$ and B is minimal this means that $(B - \{a\})^* \neq A^* \ \forall a \in B$

For example $A^* = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5\}, \{6\}\}$

Let B = {Age, Height, Gender}, C = {Age, Gender}.

We see that:

- $B^* = \{\{1\},\{2\},\{3\}, \{4\}, \{5\}, \{6\}\} = A^*$
- $C^* = \{\{1\},\{2\},\{3\}, \{4\}, \{5\}, \{6\}\} = A^*$

Also $(B^* - \{Height\}) = A^*$ this means that B is not minimal

Therefore, C is reduct of A because $C^* \neq A^*$ and C is minimal $(C - \{a\})^* \neq A^* \ \forall a \in C$

2.4 Approximation Space

For completely specified decision tables lower and upper approximations are defined on the basis of the Indiscernibility relation. Any finite union of elementary sets, associated with *B*, will be called a *B-definable* set. Let *X* be any subset of the set *U* of all cases. The set *X* is called a *concept* and is usually defined as the set of all cases defined by a specific value of the decision. In general, *X* is not a *B*-definable set. However, set *X* may be approximated by two *B*-definable sets; the first one is called a *B-lower approximation of X*, denoted by $\underline{B}X$ and a defined as $\{x \in U \mid [x]_B \subseteq X\}$. The second set is called a B-upper approximation of X, denoted by $\overline{B}X$ and defined as : $\{ x \in U \mid [x]_B \cap X \neq \emptyset \}$. The above shown way of computing lower and upper approximations, by constructing these approximations from singletons x, will be called the first method. The B-lower approximation of X is the greatest B-definable set, contained in X. The B-upper approximation of X is the smallest B-definable set containing X [5].

In our example:

U = {X1, X2, X3, X4, X5, X6} and A ={Age, Height, Gender, Accepted}
If B $\subseteq$ A and B = {Height} then B* = {{X1, X2, X3},{X5},{X4, X6}}

Suppose we have X = {X2, X3, X5} In this case we found:

Lower approximation $\underline{B}X$ = {X5} and Upper approximation $\overline{B}X$ ={X1, X2, X3, X5}

According to using lower and upper approximations discussed above, we can distinguish three regions in approximation space:

- The positive region POS(BX) = $\underline{B}X$
- The boundary region BND (BX) = $\overline{B}X$ - $\underline{B}X$
- The negative region NEG (BX) = U - $\overline{B}X$.

2.5 Rule Induction

For the inconsistent input data [2], [7], the rules induced from the lower approximation of the concept certainly describe the concept, so they are called certain. On the other hand, rules induced from the upper approximation of the concept describe the concept only possibly (or plausibly), so they are called possible.

For example: As a certain we can say: If (Age, old) and (Height, medium) then (accepted, no). As a possible we can say: If (Gender, Female) then (accepted, yes) with $\alpha = 0.67$, α is called a confidence factor and can be defined as the percentage of the number of elements that are in the elementary set and satisfy the concept for the rule from the total number of elements in the elementary set (upper approximation) in this example:

B = {Gender} then B* = {{X1, X5, X6}, {X2, X3, X4}}
X = {X2, X3} then P|X| = 2. Also, $\overline{B}X$ = {X2, X3, X4} then P|$\overline{B}X$| = 3

$$\alpha = \frac{P \mid X \mid}{P \mid \overline{B}X \mid} \qquad \alpha = \frac{2}{3} = 0.67$$

3 Modular Rough Decision Model (MRDM)

Is to convert the main task for more than a subtask and this process will be useful in several situations, for example, to avoid complications in cases of sophisticated tasks and that is to find solutions to difficult process, as a best solution such a task divided into a number of small tasks and for getting the final decision of the main task, we need a process to make compilation of these decisions to get in the end the best solution. Modular design used in various areas robotics and neural networks etc [4].

3.1 Elements of Modularity

As it has pointed by Ronco and Gawthhrop (1995), modular design to solving problems done by implementing these steps [5]:

- Decomposing the main problem into subtasks.
- Modular architecture organizing, taking into account the nature of each subtask.
- Communication between modules is important

Figure (1) represents the structure of proposed model, which performed through Graphical User Interface (GUI). MRDM proposed model allows its user to build work space, we called schema, to implement Modular Rough Decision model.

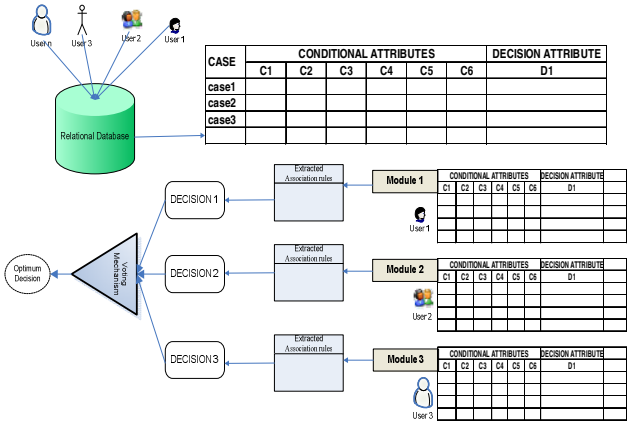


Fig. 1. Represents MRDM structure

3.2 Data Preparation

This stage concerned with using selected relational to collect the data for building the information table which, consists off column for the problem case, columns for conditions attributes and columns for decisions attributes. Data represented in rough information table in two types of attributes:

1. Conditional attributes contains data selected from the database according to the problem case, and used to take one or more decisions about this problem.

2. Decision attributes contains available decisions from the database according to transactional data.

The main objective of MRDM proposed module, is taking decision through given rough information system, by creating some modules of the main information system, then we can take decision through each module, after this step we use gating technique for taking the final decision among decisions of deferent modules. Algorithm No. 1 is to represent how to use (MRDM) proposed model for take a decision from rough information system.

Algorithm 1. Return a decision from Rough set according to gevin rule

```
  Define
  Totalrule, k , Totaldecision, poss, alfa, l
  Rule
  Descision[], C[], Avg[]
  Read nofdecisions
  {
  Input Rule
   totalrule = rough.Tables(0).columns(rule).Rows.count
          If totalrule = 0 Then
              totalrule = 0.01
        End If
    For I = 0 to nofdecisions
      Input Descision[i]
           C[i]=
rough.Tables(0).columns(rule).columns(Descision[i]).Rows.
count
          Avg[i] = C[i] / total
              k = i
          Next
          For s = 1 To k
             If Avg[s] < Avg[s-1] Then
                  Avg[s] = avg_arr[s - 1]
                  C[s] = C[s - 1]
                  Descision[s] = Descision[s - 1]
              End If
              l = s
          Next
         Totaldecision=
rough.Tables(0).columns(Descision[i]).Rows.count
          poss = C[l] / Totaldecision
          alfa = C[l] / Totalrule
       return Descision[l]
      return poss
      return alfa  }
```

Step 1: define inference rules which are given to take a decision.
Step 2: which is done by (MRDM) proposed model is determining upper and lower approximation for all possible cases to choose the optimum decision.
Step 3: calculating α degree for the different decisions according to the given rules.
Step 4: chose the optimum decision which has the greatest α degree.

Now the previous algorithm will be applied on experimental data as example to show the deference between using rough information systems and using modularity approach for dealing with rough information systems. Information collected in excel file as in table (2) and throw MRDM proposed model the data in this file have been cached and arranged into some sort of attributes.

The data represented in rough information table as previous in two types of attributes (Conditional attributes and Decision attributes).

In our example given rules is Headache = 'Yes' and Temperature > = 38
The possible decisions Flue = {'Yes', 'No'} and B = {Headache, Temperature}

Table 2. Information system for the given case study

Case	Temperature	Hypertension	Headache	Cough	Flue
1	39	120	Yes	Yes	Yes
2	42	180	Yes	No	Yes
3	39	130	No	No	No
4	38	200	Yes	Yes	Yes
5	37	170	Yes	No	No
6	37	180	No	Yes	No
7	40	190	Yes	No	No
8	40	200	Yes	Yes	Yes
9	38	200	Yes	Yes	Yes
10	37	170	Yes	No	No
11	37	180	No	Yes	Yes
12	37	120	No	No	No
13	42	130	Yes	Yes	Yes
14	37	220	Yes	No	No
15	41	180	Yes	No	No
16	39	130	No	Yes	Yes
17	40	200	Yes	Yes	Yes
18	38	130	No	No	No
19	42	220	Yes	Yes	Yes
20	37	120	Yes	Yes	Yes

B* = {{1, 2, 4, 7, 8, 9, 13, 15, 17, 19}, {3, 16}, {5, 10, 14, 20}, {6, 11, 12}}
X = {x | if headache = 'Yes' and temperature >= 38 then flue = 'Yes'}
X = {1, 2, 4, 8, 9, 13, 17, 19}
$\overline{BX}$ = {1, 2, 4, 7, 8, 9, 13, 15, 17, 19} then P|$\overline{BX}$| = 10 and P|X| = 8

$$\alpha = \frac{P \mid X \mid}{P \mid \overline{B}X \mid} \qquad \alpha = \frac{8}{10} \qquad \alpha = 0.8 \tag{1}$$

Y = {x | if headache = 'Yes' and temperature >= 38 then flue = 'No'}
Y = {7, 15} then P|Y| = 2

$$\alpha = \frac{P \mid Y \mid}{P \mid \overline{B}X \mid} \qquad \alpha = \frac{2}{10} \qquad \alpha = 0.2 \tag{2}$$

From 1, 2 then α (X) > α (Y) and the optimum decision is Flue ='Yes' with α= 0.8.

3.3 Grid Modular

This approach is a modularity technique; the main objective of using is to overcome the problem of complexity as it splits the rough decision table to smaller decision tables, which simplify reduction process by decreasing the number of attributes. This approach depends on one of two mechanisms for splitting the main information system to sub information systems (modules) . These mechanisms are:

1. Serial: according to this mechanism, the main information system splits to a given number of modules, according to the order of its attributes.
2. Random: according to this mechanism, each module has some attributes, each one is randomly chosen from the main information system.

Using of MRDM proposed model to implement grid modular approach with the same example above, the first step is to determine the number of modules, the user need to create from the main information system. Through MRDM proposed model, the number of attributes for each module will be determined. After that the user chooses the mechanism to split the main information system.

Algorithm 2. for substitute Main rough model by sub rough modules

```
Define
Total, n, d, d2, case
R() rough.tables(0).rowtype
Read nofmodule
{
        total = rough.Tables(0).Rows.count
        n = int(total / nofmodule)
        d = total / nofmodule
        d2 = (d - n) * nofmodule

            For i = 1 To nofmodule
                If i = nofmodule Then
                    n = n + d2
                End If
                For j = 1 To n
10:                 case = Int((total * Rnd()) + 1)
```

```
                    If
Rough.tables(i).Rows(case).Rows.count > 0 Then GOTO 10
                    Else
      R(case)= rough.Tables(0).Rows(case)
       Rough.tables(i).Rows(case)= R(case)
      rough.Tables(0).Rows(case).clear
                    End if
                  Next
               Next
       }
```

Step 1: determining the number of module needed to create from the main module, the user of MRDM do this step.
Step 2: calculate number of cases in each module (N cases)
Step 3: select one case randomly from the main information system and insert it into module 1, this step is repeated until it is full of the module.
Step 4: the previous step is repeated as the number of modules determined in step 1.

After creating modules from the main rough information system, MRDM proposed model allows user to define rules, which are needed to get a decision. Note that the same example used in taking decision from main rough information system represented in table (2), is used to represent using MRDM proposed model to take a decision through modularity, Tables (3,4,5 and 6) represent the four modules for the main information system of the given case study, and our given rule is the same rule used above. Headache = 'Yes' and Temperature > = 38

Table 3. Module 1 of the main information system

Case	Temperature	Hypertension	Headache	Cough	Flue
15	No	Yes	180	41	No
11	Yes	No	180	37	Yes
12	No	No	120	37	No
6	Yes	No	180	37	No
7	No	Yes	190	40	No

Table 4. Module 2 of the main information system

Case	Temperature	Hypertension	Headache	Cough	Flue
16	Yes	No	130	39	Yes
1	Yes	Yes	120	39	Yes
17	Yes	Yes	200	40	Yes
9	Yes	Yes	200	38	Yes
18	No	No	130	38	No

Table 5. Module 3 of the main information system

Case	Temperature	Hypertension	Headache	Cough	Flue
8	Yes	Yes	200	40	Yes
20	Yes	Yes	120	37	Yes
2	No	Yes	180	42	Yes
19	Yes	Yes	220	42	Yes
10	No	Yes	170	37	No

Table 6. Module 4 of the main information system

Case	Temperature	Hypertension	Headache	Cough	Flue
13	Yes	Yes	130	42	Yes
5	No	Yes	170	37	No
14	No	Yes	220	37	No
3	No	No	130	39	No
4	Yes	Yes	200	38	Yes

After defining rules, one decision is taken from each module with α degree. Final step in taking decision through MRDM proposed model using modularity approach is gating process. In MRDM proposed model voting technique is used as a gating process, this done by making vote between the decisions taken by the modules. The voting process is taking into account two factors which are α degree and possibility degree. Possibility degree is calculated in MRDM proposed model as percentage between the numbers of cases achieve the given rule to number of cases achieve chosen decision.

Algorithm 3. Gating a decision from N Rough modules

```
  Define
  K, netdecision, netalfa
  Descision[],     C[],     alfa[],     poss[],avgalfa[],
avgposs[],decval[]
  Read nofdecisions, nofmodules
  {
  For I = 0 to nofdecisions
  Input Descision[i]
           For N = 1 to nofmodules
      If rough.Tables(N).getdecision(decision) =
Descision[i] then
           {
      C[i] = C[i] + 1
      Alfa[i]=alfa[i] + rough.Tables(N).getdecision(alfa)
```

```
                    Poss[i] = poss[i] +
rough.Tables(N).getdecision(poss)
                }
        End if
              Next
              Avgalfa[i] = alfa[i]/C[i]
              Avgposs[i] = poss[i]/C[i]
              Decval[i] = Avgalfa[i] * Avgposs[i]
              K = i
            Next
            For s = 1 To k
                If decval[s] < decval[s-1] Then
                   Descision[s] = Descision[s - 1]
                End If
                l = s
            Next
          Netdecision  = Descision[l]
       Netalfa = decval[l]

           return Netdecision
         return netalfa
   }
```

Step 1: Using Algorithm No. 1 to get a decision from each module

Step 2: Calculate α degree and possibility degree for the optimum decision of each module as following:

$\alpha = \frac{P \mid X \mid}{P \mid \overline{B}X \mid}$ where X = {x | number of cases achieve the optimum decision according to the given rule} and $\overline{B}X$ is the upper approximation of the given rules

Possibility = $\frac{P \mid X \mid}{P \mid D \mid}$ where D = {x | number of cases achieve the optimum decision}

Step 3: apply voting technique to select the optimum decision of the model among the decisions achieved in each module, according to α degree and Possibility degree

The vote of each decision is calculated as a summation of [α * possibility] for each module achieve the decision, figure (3) explains how to implement the voting process.

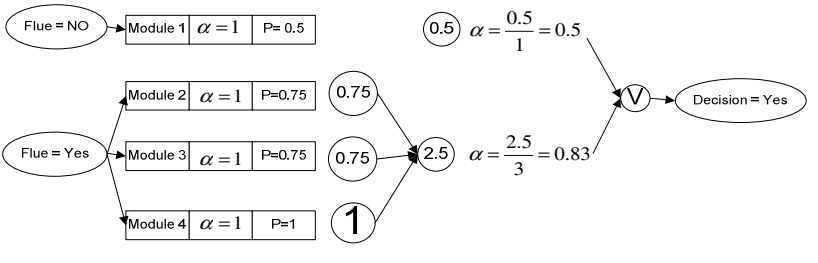


Fig. 3. Voting process implementation

3.4 Advantages of Modularity

A list of advantages of modularity is given below:

- Modularity may imply reduction of number of parameters, which will allow and increase in computing speed and better generalization capabilities.
- System helping to understand the role that each task plays within the complete system.
- Computation needed in each subsystem design are typically lower, as we need *N Permutation K* tries to search in rough information system as: N = K * L
  K = number of attributes in the given rule and L = number of cases

Proof
If we have main rough information system for L cases and we need to take a decision according to rule contains K attributes then we have T= N *Permutation K* trials

$$T = N * (N - 1) \ldots (N - K\text{-}1) \tag{1}$$

If we need to split the main rough information system to M number of modules
Then we have $\mathcal{T} = {}_{P\ K}^{N/M}$ trials for each model

$$\mathcal{T} = {}_{P\ K}^{N/M} * ({}_{P\ K}^{N/M} - 1) \ldots ({}_{P\ K}^{N/M} - k\text{-}1) \tag{2}$$

If we multiply equation (2) by M

$$M\mathcal{T} = M * {}_{P\ K}^{N/M} * ({}_{P\ K}^{N/M} - 1) \ldots ({}_{P\ K}^{N/M} - k\text{-}1)$$

$$\mathcal{T} = N * (N - M) \ldots (N - M(k\text{-}1)) \tag{3}$$

From (1), (2) and (3) Then $M\mathcal{T} < T$

This means that total number trials to search in all modules smaller than number of trials to search in main rough model. So, computations needed to take a decision through splitting the main rough information system, are typically lower than those needed to take a decision directly from the main rough information system.

4 Conclusion and Future Work

Modular approach reduces computation complexity. In many cases, appropriate decomposition of modules is related to variety of users. The idea is to ignore interconnection among subsystems in the design stage. Design effort and computation needed in each subsystem design are typically lower. The system will also be easier to debug and maintain. In the next phase, the goal is to add another level of modularity by applying fuzzy roles on the data in modules.

References

1. Degang, C., Suyun, Z.: Local reduction of decision system with fuzzy rough sets. Fuzzy Sets and Systems 161, 1871–1883 (2010)
2. Thangavel, K., Pethalakshmi, A.: Dimensionality reduction based on rough set theory. Applied Soft Computing 9, 1–12 (2009)
3. Qian, J., Liang, D., Li, Z.H., Dang, C.: Measures for evaluating the decision performance of a decision table in rough set theory. Information Sciences 178, 181–202 (2008)
4. Tseng, B.: Modular neural networks with applications to pattern profiling problems. Neurocomputing (2008)
5. Melin, P., Gonzalez, C., Bravo, D., Gonzalez, F., Martinez, G.: Modular Neural Networks and Fuzzy Sugeno Integral for Pattern Recognition. STUDFUZZ, vol. 208, pp. 311–326 (2007)
6. Grzymala, J.W., Siddhaye, S.: Rough Set Approaches to Rule Induction from Incomplete Data. In: The 10thInternational Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, Perugia, Italy, July 4-9, vol. 2, pp. 923–930 (2004)
7. Pedryez, W., Gomide, F.: An Introduction to Fuzzy Sets Analysis and Design. Massaachusetts Institute of Technology (1998)
8. Lefteri, H.T., Robert, E.U.: Fuzzy and Neural Approaches in Engineering. A Wiley-Interscience Publication (1997)
9. Zadeh, L.: Fuzzy Sets. Information and Control 8, 338–353 (1965)

Cost Optimized Approach to Random Numbers in Cellular Automata

Arnab Mitra[1,3] and Anirban Kundu[2,3]

[1] Adamas Institute of Technology, West Bengal - 700126, India
mitra.arnab@gmail.com
[2] Kuang-Chi Institute of Advanced Technology, Shenzhen - 518057, P.R. China
anirban.kundu@kuang-chi.org, anik76in@gmail.com
[3] Innovation Research Lab (IRL), West Bengal - 711103, India

Abstract. In this research work, we are trying to put emphasis on on the cost effective generation approach of high quality random numbers using one dimensional cellular automaton. Maximum length cellular automata with higher cell number, has already been established for the generation of highest quality of pseudo random numbers. Sequence of randomness quality has been improved using DIEHARD tests reflecting the varying linear complexity compared to the maximum length cellular automata. The mathematical approach for proposed methodology over the existing maximum length cellular automata emphasizes on flexibility and cost efficient generation procedure of pseudo random pattern sets.

Keywords: Random Pattern Generator, Pseudo Random Number Generator (PRNG), Randomness Testing, Prohibited Pattern Set (PPS), Cellular Automata (CA).

1 Introduction

Random numbers [1] play an important role in present days scientific and engineering applications. The application field includes the area varying from statistics or mathematics to VLSI circuit testing [2] or game playing. By properties**,** random numbers occur in a progression such that the values must be uniformly distributed over a defined interval and it must be impossible to guess future values based on past or present ones. Pseudo random numbers are generated by means of executing a computer program based on any particular algorithm. The adjective "pseudo", refers the predetermined manner in which "randomness" is being formed. The algorithms and computer programs that produce this type of random numbers are usually referred to as pseudo-random number generators (PRNG). In creation of random numbers, most random number generators have need of an initial number to be used as the preparatory point, known as a "seed". It is necessary to normalize the distributions over some specified margin so that each differential area is equally populated.

On the other hand, a cellular automaton (pl. cellular automata, in short it is named as CA) is a discrete mathematical model studied in the field of computability theory to complexity science and from theoretical biology to microstructure modeling. Cellular Automata consist of a regular framework of cells. Each of the cells is in a state, which is

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 609–618.
springerlink.com

either in “On or 1” or “Off or 0”. This grid of cells can be of multi-dimensions [3]. For each cell surrounded with its neighborhood, usually including the cell itself is defined with respect to the particular cell. A new generation is created with some predetermined rule, with the progress of time. A mathematical function determines the value of the new state with respect to the current state of the cell and the states of the cells in its neighborhood. The simplest nontrivial CA is one-dimensional, with only two possible values per cell. The neighbors of the cell are the nearby cells on either side of it. A cell along with its two neighbors, form a neighborhood of 3 cells. Thus there exists a total 2^3=8 possible patterns, for this 3 cell CA. There exists total 2^8=256 possible rules. Generally these 256 CAs are referred to by their Wolfram code. Wolfram code gives each rule a distinct number from 0 to 255. These 256 CAs, either individually or collectively, have already been analyzed and compared in a number of research papers. Several research works show the rule 30 and the rule 110 CAs are mostly interesting [4]. Fig. 1 represents the typical structure of 3-cell Null Boundary CA.

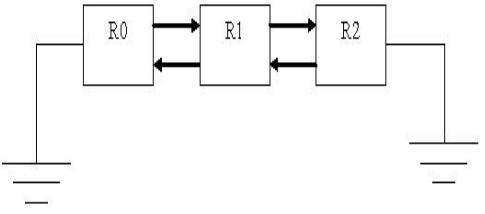


Fig. 1. Typical structure of 3-cell Null Boundary CA

Rest of the paper is organized as follows: Section 2 briefly discusses about related work; Section 3 describes the proposed work; Section 4 shows the experimental results; Section 5 draws the conclusion.

2 Related Work

Several methods have been implemented to generate good quality random numbers [5]. By past efforts [6-9], it has been established that the maximum length cellular automata exhibits the maximum quality of randomness.

The most common way to generate pseudo random numbers is to follow a combination of “randomize” and “rand” functions. Random patterns can be achieved based on the following recursive PRNG equation.

$$X_{n+1}=P_1X_n+P_2 \text{ (Mod N)} \quad (1)$$

where, P1 and P_2 are two prime numbers;
N is range of random numbers;
X_n is calculated recursively using the value of X_0 as base value.
X_0 is termed as seed and it is also a prime number.
If X_0 (seed) is same all time, then it produces pseudo random number.

In the research on maximum length cellular automata, it is well established that resulting maximum length cycle includes the maximum amount of randomness with the increased number of cells. In maximum length cellular automata, prohibited pattern set (PPS) is excluded [7] [9] [10] from the cycle for achieving higher quality of randomness [5] [11-13].

3 Proposed Work

In the generation process of pseudo random pattern using cellular automata, a cycle is responsible for yielding the pseudo random patterns. If a single cycle with larger numbers of states is used to generate the random patterns, the associated cost would be increased. All the cost associated with the generation process of random numbers; i.e., time, design and searching costs are having higher values as the complexity is directly proportional to the number of cell sizes used in the cycle. So, it is convenient to reduce the cycle without affecting the randomness quality of the generated random patterns for reducing these associated costs.

In the proposed optimized maximum length CA (one dimensional), we propose the decomposition of the cycle into more relevant sub-cycles such that our concerning complexities cost can be reduced. Fig. 2 represents the flowchart of the proposed system.

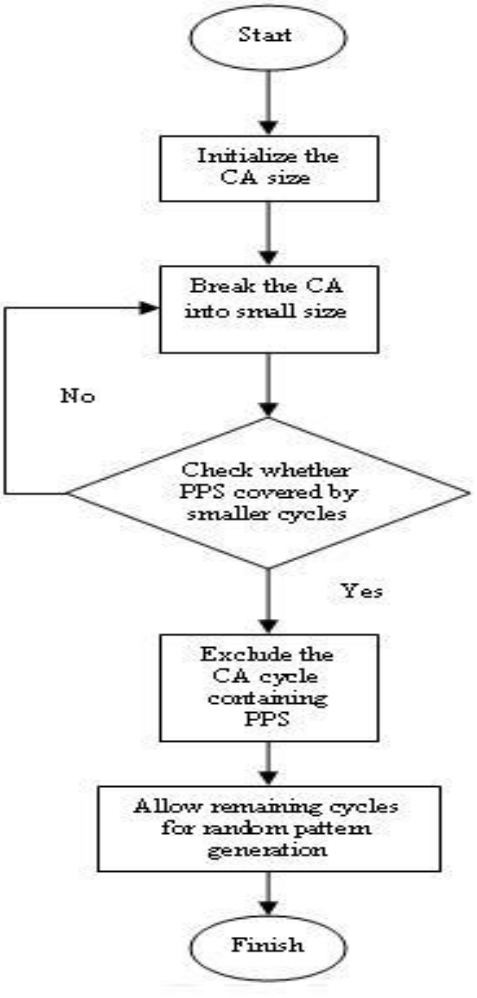


Fig. 2. Flowchart of proposed system

A new mathematical approach has been proposed according to the flowchart of the proposed system, to achieve the same amount of randomization in less cost with respect to various complexities and hardware implementation.

The cycle is divided into two or more equal sub-cycles instead of taking the full cycle of maximum length in optimized maximum length CA. These sub-cycles are capable to generate good quality random patterns as equivalent to the randomness quality achieved from maximum length CA. The following Algorithm1 has been used for the decomposition of any n-cell maximum length CA.

Algorithm 1. Cycle Decomposition

Input: CA size (n), PPS Set
Output: m-length cycles excluding PPS

Step 1: Start
Step 2: Initialize the number of n-cell CA for generating random patterns using n-cell CA
*Step 3: Decompose the cell number (n) into two equal numbers (m) such that n=2*m*
Step 4: Verify for each PPS whether the PPS belongs to a single smaller cycle CA or not
Step 5: Repeat Step 3 and Step 4 until each of the PPS belongs to separate smaller cycles
Step 6: Permit m-length cycles of n-cell CA after excluding all the PPS containing cycles
Step 7: Stop

In our approach, the main concern is to minimize the PPS within the random set of patterns. Therefore it has been taken care of that the occurrence of every PPS must be completed in some of the smaller sub-cycles, such that by eliminating all those smaller cycles, the remaining cycles can be allowed to generate random patterns. Thus, this methodology implies a better cost effectiveness approach. The proposed methodology thus simplifies the design complexity and empowers the searching complexity. The terminology design complexity refers to the implementation procedure for generation of random pattern and empowering searching complexity means the zero overhead for keeping track for PPS for random pattern generation.

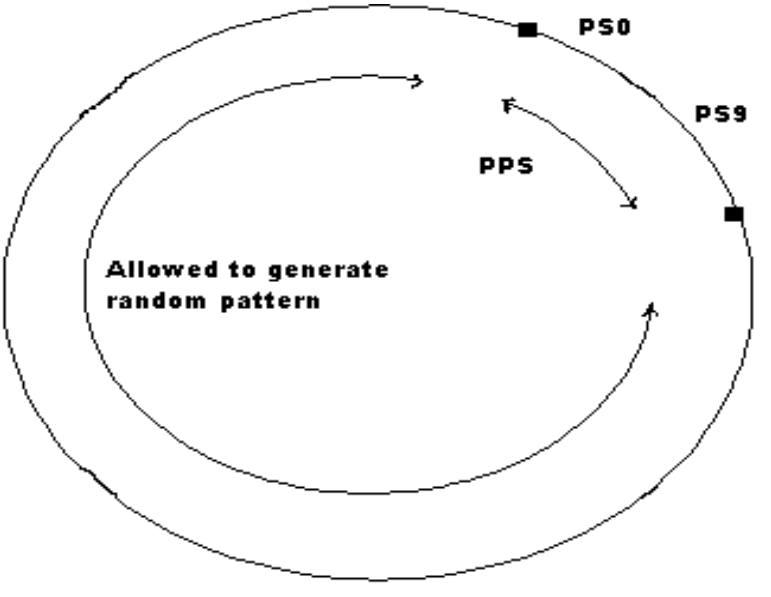
In comparison with n-cell maximum length CA, more number of smaller cycles instead of one maximum length cycle should be used.

Consider, CA size = n;
Then, $2^n = 4k$ (assume, $k=2^m$)
We have $2^n = 4(2^m)$; i.e., 4 numbers of equal length cycles.
Thus 'm' is always less than 'n'.
Let, one maximum length CA is divided into two sub-cycles.
Then, $2^n = (2^{m1} + 2^{m2})$
If m1= m2
Then it becomes $2^n = 2\ (2^{m1})$
Upon illustrating this model, let us assume the case where maximum length CA, n=8

So, $2^8=2^7+2^7$ (i.e., two equal length of CA cycles)
$=2^6+2^6+2^6+2^6$ (i.e., four equal length of CA cycles)
$=2^5+2^5+2^5+2^5+2^5+2^5+2^5+2^5$ (i.e., eight equal length of CA cycles)
$=2^4+2^4+2^4+2^4+2^4+2^4+2^4+2^4+2^4+2^4+2^4+2^4+2^4+2^4+2^4+2^4$ (i.e., sixteen equal length of CA cycles)

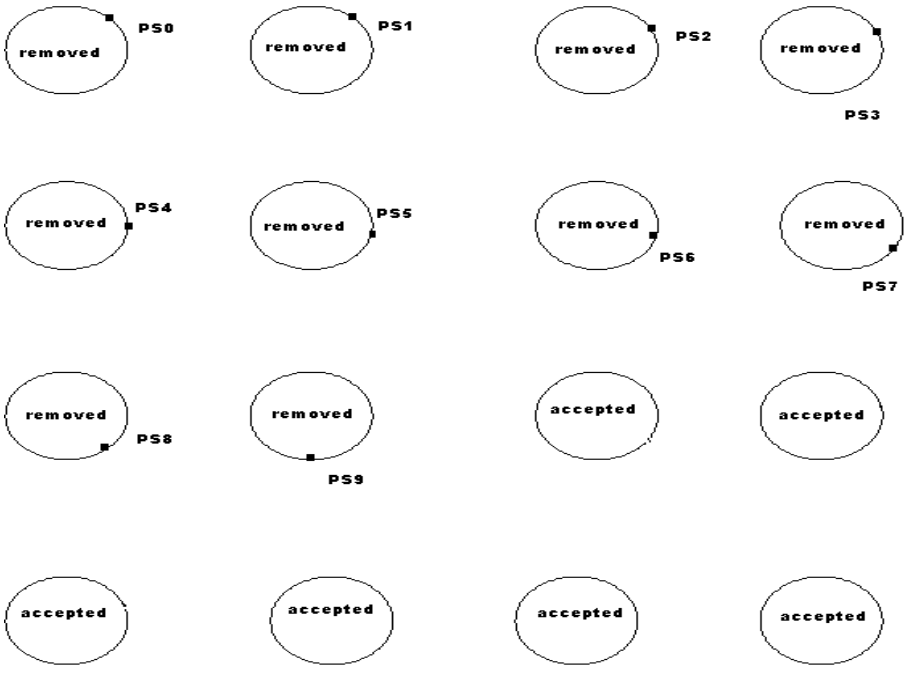
The PPS is excluded from the cycle as per the procedure for generating the maximum random pattern in maximum length CA. In our procedure, the PPS can be totally removed as we are having more number of cycles for generation of random sequences. In proposed methodology, the cycles producing PPS can be removed from the generation of pseudo random numbers.

According to the proposed methodology, a CA size of n=24 might be decomposed into equal length smaller cycles instead of one maximum length circle. In this case it can be divided into 16 smaller cycles of length 2^{20}. Let us assume that there exists ten numbers of prohibited patterns as PPS. So in worst case, assuming every single prohibited pattern occurs in a single cycle, there exists (16 -10)=6 number of CA cycles having 2^{20} cycle length for generation of random patterns. The pattern generation on this scenario is followed as in Fig. 3. Fig. 3(a) shows one maximum length cycle with prohibited patterns, and on the other hand, Fig. 3(b) shows 16 equal length smaller cycles where some of the cycles contain prohibited patterns only. The following figure is based on Null Boundary 24-cell CA. The PPS is denoted as $\{PS_0, PS1,\ldots\ldots PS_9\}$.



(a)

Fig. 3. (a) Maximum length CA Cycle for n=24; and, (b) Proposed equal length CA of smaller cycle size



(b)

Fig. 3. (*continued*)

4 Experimental Observations and Result Analysis

With the help of p-value analysis, generated for each sample data set by Diehard battery series test, a decision is made whether the test data set passes or fails the diehard test. Diehard returns the p-value, which should be uniform [0, 1) if the input file contains strictly independent random bits. These p-values are obtained by a function, p=F(x), where 'F' is implicit distribution of the random sample variable 'x'. $p<0.025$ or $p>0.975$ means the RNG has "failed the test at the 0.05 level" [5].

In terms of the total number of pass/fail of the Diehard test cases, the comparison Table 1 shows that the proposed methodology is equivalent to maximum length CA.

Table 1. Performance result through Diehard Test

Diehard Test Number	Name of the test	Max-length CA		Proposed methodology	
		n=23	n=64	n=23	n=64
1	Birthday Spacing	Pass	Pass	Pass	Pass
2	Overlapping Permutations	Pass	Pass	Pass	Pass
3	Ranks of 31x31 and 32x32 matrices	Pass	Pass	Pass	Pass
4	Ranks of 6x8 Matrices	Pass	Pass	Pass	Pass
5	The Bit stream Test	Fail	Fail	Fail	Fail
6	Monkey Tests OPSO,OQSO, DNA	Fail	Pass	Fail	Pass
7	Count the 1`s in a Stream of Bytes	Pass	Pass	Pass	Pass
8	Count the 1' s in Specific Bytes	Fail	Pass	Fail	Pass
9	Parking Lot Test	Pass	Pass	Pass	Pass
10	Minimum Distance Test	Pass	Pass	Pass	Pass
11	The 3DSpheres Test	Pass	Pass	Pass	Pass
12	The Sqeeze Test	Fail	Pass	Fail	Pass
13	Overlapping Sums Test	Fail	Pass	Fail	Pass
14	Runs Test	Pass	Pass	Pass	Pass
15	The Craps Test	Pass	Pass	Pass	Pass
Total Number Passes		10	14	10	14

With reference to the result obtained in the Table 1, the quality of randomness can be represented in a bar chart. The Fig. 4 and Fig. 5 show the quality achieved for both of the procedures under test.

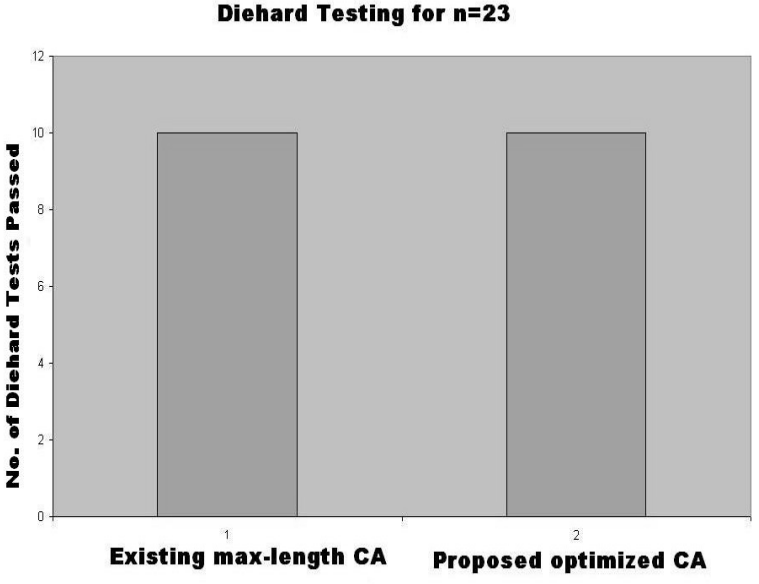


Fig. 4. Randomness comparison for n=23 cell CA

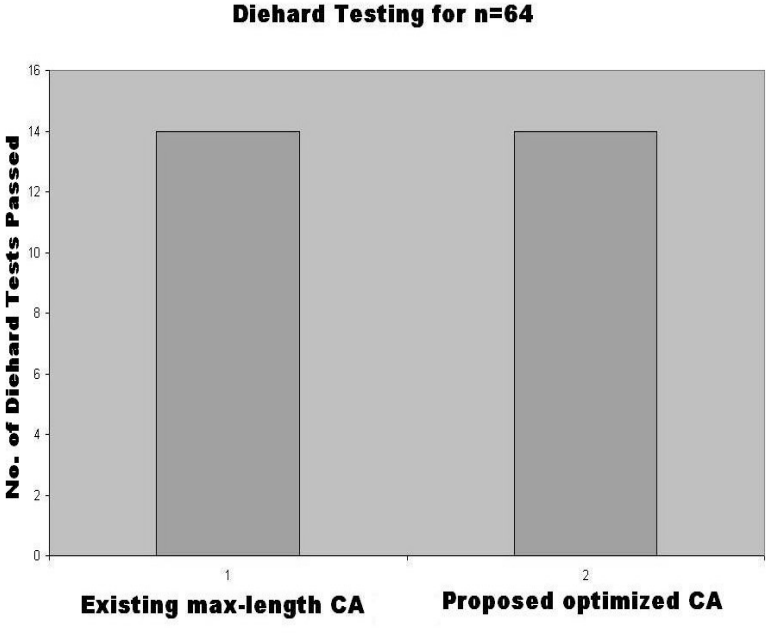


Fig. 5. Randomness comparison for n=64 cell CA

Table 2. Complexity Comparison between Maximum Length CA and Proposed Methodology

Name of the Complexity	Comparison Result
Space	Same
Time	Slightly Improved
Design	Improved
Searching	Improved

Table 2 signifies that the space complexity is same for both procedures as total length of an n-cell CA is same for both the cases, but there are some changes in other complexities. Other complexities have been improved in case of our proposed methodology. The proposed methodology is allowed only to generate random patterns from smaller cycles that exclude PPS. The PPS exclusion feature from the main cycle, which is responsible for generating random patterns, improves the design and searching complexities.

The quality of random pattern generation through proposed methodology should be same as it was in case of maximum length CA. Thus it is a major advantage of our proposed methodology that it has higher flexibility along with less amount of cost involved. More number of smaller cycles generates the same result of randomness following our technique with an added advantage of no prohibited pattern set.

5 Conclusion

Based on the result analysis from Table 1 and Fig. 4 & Fig. 5, the conclusion can be made that the proposed methodology has achieved the same degree and quality of randomness as compared to maximum length CA. Hence, the proposed methodology is suitable for generating random sequences as the cost associated is much cheaper in terms of time complexity and hardware implementation. The proposed method uses only the sub-cycles which consist of lesser number of states than of maximum length CA. Thus it completes its full cycle in lesser time and requires lesser number of hardware implementation cost. The time complexity of the proposed method increases with the increased number of cells in CA. It is also convenient that the proposed RNG is more flexible as it completely excludes the prohibited pattern set (PPS).

References

[1] Wolfram Mathematica Tutorial Collection: Random Number Generation, http://www.wolfram.com/learningcenter/tutorialcollection/RandomNumberGeneration/RandomnumberGeneration.pdf

[2] Das, S., Dey, D., Sen, S., Sikdar, B.K., Chaudhuri, P.P.: An efficient design of non-linear CA based PRPG for VLSI circuit testing. In: ASP-DAC, Japan (2004)

[3] Das, S., Sikdar, B.K., Pal Chaudhuri, P.: Characterization of Reachable/Nonreachable Cellular Automata States. In: Sloot, P.M.A., Chopard, B., Hoekstra, A.G. (eds.) ACRI 2004. LNCS, vol. 3305, pp. 813–822. Springer, Heidelberg (2004)

[4] Das, S., Sikdar, B.K.: Classification of *CA* Rules Targeting Synthesis of Reversible Cellular Automata. In: El Yacoubi, S., Chopard, B., Bandini, S. (eds.) ACRI 2006. LNCS, vol. 4173, pp. 68–77. Springer, Heidelberg (2006)
[5] Robert, G.B.: Dieharder: A Random Number Test Suite, C program archive dieharder, version 1.4.24 (2006), `http://www.phy.duke.edu/~rgb/General/dieharder.php`
[6] Das, S., Rahaman, H., Sikdar, B.K.: Cost Optimal Design of Nonlinear CA Based PRPG for Test Applications. In: IEEE 14th Asian Test Symposium, Kolkata (2005)
[7] Das, S., Kundu, A., Sikdar, B.K., Chaudhuri, P.P.: Design of Nonlinear CA Based TPG Without Prohibited Pattern Set In Linear Time. Journal of Electronic Testing: Theory and Applications (2005)
[8] Das, S., Kundu, A., Sikdar, B.K.: Nonlinear CA Based Design of Test Set Generator Targeting Pseudo-Random Pattern Resistant Faults. In: Asian Test Symposium, Taiwan (2004)
[9] Das, S., Kundu, A., Sen, S., Sikdar, B.K., Chaudhuri, P.P.: Non-Linear Celluar Automata Based PRPG Design (Without Prohibited Pattern Set) In Linear Time Complexity. In: Asian Test Symposium, China (2003)
[10] Ganguly, N., Nandi, A., Das, S., Sikdar, B.K., Chaudhuri, P.P.: An Evolutionary Strategy To Design An On-Chip Test Pattern Generator Without Prohibited Pattern Set (PPS). In: Asian Test Symposium, Guam (2002)
[11] Sikdar, B.K., Das, S., Roy, S., Ganguly, N., Das, D.K.: Cellular Automata Based Test Structures with Logic Folding. In: VLSI Design, India (2005)
[12] Das, S., Sikdar, B.K., Chaudhuri, P.P.: Nonlinear CA Based Scalable Design of On-Chip TPG for Multiple Cores. In: Asian Test Symposium, Taiwan (2004)
[13] Das, S., Ganguly, N., Sikdar, B.K., Chaudhuri, P.P.: Design of A Universal BIST (UBIST) Structure. In: VLSI Design, India (2003)

Selection of Views for Materializing in Data Warehouse Using MOSA and AMOSA

Rajib Goswami, D.K. Bhattacharyya, and Malayananda Dutta

Department of Computer Science & Engineering
Tezpur University
Tezpur, India
{rgos,dkb,malay}@tezu.ernet.in

Abstract. By saving or materializing a set of derived relations or intermediate results from base relations of a data warehouse, the query processing can be made more efficient. It avoids repeated generation of these temporary views while generating the query responses. But as in case of a data warehouse there may be large number of queries containing even larger number of views inside each query, it is not possible to save each and every query due to constraint of space and maintenance costs. Therefore, an optimum set of views are to be selected for materialization and hence there is the need of a good technique for selecting views for materialization. Several approaches have been made so far to achieve a good solution to this problem. In this paper an attempt has been made to solve this problem by using Multi Objective Simulated Annealing(MOSA) and Archived Multi-Objective Simulated Annealing(AMOSA) algorithm.

Keywords: Data Warehouse, View Materialization, View Selection, Multi-Objective Optimization, Simulated Annealing, Multi-Objective Simulated Annealing (MOSA), Archived Multi-Objective Simulated Annealing (AMOSA).

1 Introduction

In conventional database management system concept, a view is defined as a derived relation on some base relations. A view defines a function from a set of base tables to a derived table. If these views are materialized by storing the tuples of the views in the database, then the query processing becomes much faster as there is no need of re-computing the views while processing the queries. But it is not possible to save all these views, as we may have to consider a large number of frequent queries and the related temporary views in limited availability of space. Again the process of updating a materialized view, known as view maintenance, in response to changes in the base data is also involved. Therefore, there is a need for selecting an appropriate set of views to materialize for a set of queries for increasing query performance in terms of query processing cost and view maintenance cost. This is known as materialized view selection problem [3, 4, 7]. The research work on view selection for materializing in data warehouses started in early 90s with some heuristic greedy algorithms [3, 4, 7, 9-16].

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 619–628.
springerlink.com

When the dimensions of data warehouses grow, the solution space of this optimization problem also grows exponentially. Thus the problem becomes NP-hard [1-4, 7, 8]. And therefore, most of the recent approaches use either randomized algorithms like Simulated Annealing(SA), Genetic Algorithm, Parallel Simulated Annealing(PSA), particle swarm optimization(PSO) and memetic algorithm(MA) [5, 6, 17-22] or apply data mining techniques[23-25], to deal with materialized view selection problem. A Directed Acyclic Graph (DAG) known as AND-OR View Graph [7, 8] is commonly used to represent the relationships among query, view and base tables while designing the input data structure for view selection for materialization algorithms. Another means of representation used in View Selection for materializing in data warehouse is by using a DAG representing a set of frequently asked queries by a query processing strategy of warehouse views which is termed as Multiple View Processing Plan (MVPP) [17]. In most recent approaches, this problem is represented as a single objective optimization problem. Here instead of representing the problem as single objective optimization problem, we are formulating the problem for optimization of multiple objectives such as total query cost and total view maintenance cost of a view processing plan of a set of frequent queries. We have used Multi Objective Simulated Annealing (MOSA) [27, 28] and another recent Multi Objective Simulated Annealing technique termed as Archived Multi-Objective Simulated Annealing(AMOSA) [29] for handling the problem and a comprehensive analysis is presented.

2 Backgrounds

2.1 Multiple View Processing Plan (MVPP) and Selecting Views to Be Materialized [17]

To select a set of views for materializing for efficient and cost effective query processing, the two basic inputs we have are: (i) a set of frequent global queries with their access frequencies and (ii) a set of base tables with their maintenance frequencies. In [17], a DAG termed as Multiple View Processing Plan (MVPP) is defined representing a query processing strategy of data warehouse views. Here the root nodes correspond to data warehouse queries and leaf nodes correspond to base relations by connecting intermediate query processing results and final query results as intermediate vertices. The view selection problem is to select some of the intermediate vertices for materializing, so that the total query processing cost and total updating cost of the materialized views become minimum where total numbers of vertices or temporary views that can be materialized depend on specific amount of available space.

Implementation of two Multi-Objective Simulated Annealing algorithm based techniques to deal with view selection for materializing in data warehouse problem by using MVPP representaion is presented in this paper. In the subsequent subsections, MOSA and AMOSA algorithms for multi-objective optimization have been discussed.

2.2 Multi-Objective Simulated Annealing (MOSA)

Multi-Objective Optimization is defined as the optimization of D simultaneous objective functions:

$$y_i=f_i(\mathbf{x}), \quad \text{where } i=1,\ldots\ldots, D, \tag{1}$$

while satisfying constraints, if any, and $\mathbf{x}=(x_1, x_2, \ldots\ldots, x_P)$ are P decision variables of the objective function[27, 28].

Thus in case of a Multi-Objective Optimization problem, for minimization, the problem may be expressed as:

$$\text{Minimize} \quad \mathbf{y}=\mathbf{f}(\mathbf{x}) \equiv (f_1(\mathrm{x}), f_2(\mathrm{x}),\ldots\ldots f_D(\mathrm{x})) \tag{2}$$

under the other constraints specified, if any. For two solutions x_i and x_j, if the optimization is a minimization problem, then if

$$\begin{aligned} &f_k(x_i) \leq f_k(x_j) \quad \forall k \in 1, 2, \ldots, D \quad \text{and} \\ &\exists k \in 1, 2, \ldots, D, \text{ such that } \quad f_k(x_i) < f_k(x_j) \end{aligned} \tag{3}$$

then x_i is said to dominate x_j, which is written as $x_i \prec x_j$. Thus by slight change in notations we can express that for two solutions **a** and **b**, $\mathbf{a} \prec \mathbf{b}$ iff $f(\mathbf{a}) \prec f(\mathbf{b})$. A set of solutions $\boldsymbol{P}$ is called a *non-dominating set of solutions*, when

$$\mathbf{a} \not\prec \mathbf{b}, \mathbf{b} \not\prec \mathbf{a} \; \forall \mathbf{a}, \mathbf{b} \in \boldsymbol{P} \tag{4}$$

If a solution is globally non-dominated i.e. there is no feasible solution available that dominates this solution, then this solution is called *Pareto-optimal solution*. The set of Pareto-optimal solutions is termed as *Pareto front.*

Multi-Objective Simulated Annealing (MOSA) Algorithm: In MOSA, Kevin I. Smith et al. in [27], proposed to use an energy definition in terms of the current estimate of the Pareto front, F , the set of mutually non-dominating solutions found thus far in the annealing. Here, for current solution x, new solution x′ , and current estimate of the Pareto front F, F is defined as:

$$\mathrm{F}= F \cup \{\mathrm{x}\} \cup \{\mathrm{x}'\}, \tag{5}$$

Now let F_x be the elements of F*s* that dominate x and $\mathrm{F}_{x'}$ be the elements of F that dominate x′ i.e.

$$\mathrm{F}_x=\{\mathrm{y} \in \mathrm{F} \mid \mathrm{y} \prec \mathrm{x}\} \tag{6}$$

If $|\mathrm{F}_x|$, $|\mathrm{F}_{x'}|$ and $|\mathrm{F}|$ are the number of elements in respective sets; the energy difference $\delta E(\mathrm{x}', \mathrm{x})$ between the proposed and the current solution is defined as

$$\delta E(x', x) = \frac{1}{|\overline{F}|}(|\overline{F_{x'}}| - |\overline{F_x}|) \tag{7}$$

Unlike uni-objective simulated annealing, MOSA yields a set of mutually no dominating solutions which is only an approximation to the *true Pareto front* [27, 28].

2.3 Archived Multi-Objective Simulated Annealing (AMOSA)

In Pareto-domination-based MOSAs as discussed in sub-section 2.2., the acceptance criterion between the current and a new solution is expressed in terms of the difference in the number of solutions that they dominate[27-29] but not by the amount of

domination that takes place. In [29], Bandyopadhyay et al., proposed a new MOSA referred as Archived Multi-objective Simulated Annealing (AMOSA). The AMOSA incorporates the concept of amount of dominance in order to determine the acceptance of a new solution. Here for two solutions x and x', amount of domination by x on x' is defined as:

$$\Delta dom_{x,x'} = \prod_{i=1, f_i(x) \neq f_i(x')}^{D} (| f_i(x) - f_i(x') | / R_i) \tag{8}$$

Where D=number of objectives, R_i is the range of the i^{th} objective.

In case R_i is not known a priori, whatever solutions present so far (including the current solution and the new proposal) are considered for finding the range R_i.

3 The View Selection for Materializing in Data Warehouses as Multi-Objective Optimization Problem

The view selection for materializing from an MVPP as discussed in sub-section 2.1 and using the notations as reported in [5, 6, 17], is to determine a set of vertices M from the set of vertices V, i.e. $M \subseteq V$ of an MVPP, such that $\forall v \in M$, if $T(v)$ is materialized then the total cost of query processing and total cost of materialized view maintenance is minimum.

Using the notions and notations used in[17], if V is the set of vertices of MVPP G, and M is the set of views or vertices that are materialized, i.e. $M \subseteq V$, then under the constraint $\Sigma_{v \in M} A_v \leq A$, where A_v denotes space required for materializing view v and A be the total space available for materializing or storing the views; the view selection problem is to minimize:

$$\text{Query cost, } Q_G(M) = \sum_{q \in R} f_q(C_a^q(v)) \tag{9}$$

and view maintenance or updating cost,

$$U_G(M) = \sum_{m \in M} f_m(C_u^m(r)) \tag{10}$$

Where,

R is the set of root nodes of MVPP, G,
f_q is the access frequency of query $q \in R$,
$C_a^q(v)$ is the processing cost of query q, $q \in R$, by accessing vertices v, where $v \in V$, when M, $M \subseteq V$, is materialized,
f_m is the updating or maintenance frequency of materialized views $m \in M$ and $C_u^m(r)$ is the cost of updating materialized views $m \in M$ by accessing or updating the vertices r where $r \in V$.

Thus the multi-objective optimization problem may be defined as:

$$\text{Minimize } \mathbf{y}=\mathbf{f}(M) \equiv (Q_G(M), U_G(M)) \tag{11}$$

under the constraint $\Sigma_{v \in M} A_v \leq A$, where A_v denotes space required for materializing view v and A be the total space available for materializing.

For solutions S_0 and S_1 of (11), under the constraint $\Sigma_{v \in M} A_v \leq A$,

$S_0 \prec S_1$,
iff ,
($Q_G(S_0) \leq Q_G(S_1)$ and $U_G(S_0) \leq U_G(S_1)$)
and
($Q_G(S_0) < Q_G(S_1)$ or $U_G(S_0) < U_G(S_1)$) (12)

That is, if the logical condition (12) is satisfied by the solutions S_0 and S_1 for the problem (11) then S_1 is said to be dominated by S_0 and expressed as $S_0 \prec S_1$. If $S_0 \not\prec S_1$ and $S_1 \not\prec S_0$ then S_0 and S_1 are said to be non-dominating solutions. Here our objective is to find a set of non-dominating solutions of this problem which is an approximation to the true Pareto-front.

4 View Selections Using Multi-Objective Simulated Annealing

4.1 Solution Representation

Our problem is to find an archive of non-dominating solution, by knowing an MVPP, say *G,* for minimizing query cost and updating cost of an MVPP with the constraint of space. To represent the solution, the nodes of the MVPP are labeled in a specific order starting at the base relations of the graph [5, 6, 17]. All intermediate nodes, (i.e., intermediate temporary views) of the MVPP graph are kept in an array. Thus for *m* number of intermediate nodes or intermediate views, the views are indexed as: v_i , *i=0* to *m-1*. Then in our solution representation, a solution string *S* is represented such that if *i*-th view v_i is selected for materialization, then *i*-th character of *S* is set to "1" and if *i*-th view v_i is not selected then *i*-th character of *S* is set or represented with "0". For example, if a solution string looks like "10000101" for an MVPP, then it means that the MVPP graph has eight number of intermediate nodes and out of them, first, sixth and eighth views are selected.

For computing space requirement for selected views, query processing cost of the MVPP, and for computing updating cost or maintenance cost of the materialized views of the MVPP, (i) the number of rows (or records) by each relation corresponding to the nodes of the MVPP are kept in an array, and (ii) a separate array is maintained for keeping query frequencies of different queries. In our approach, for programming convenience, updating frequency of base tables is assumed to be fixed for an MVPP.

4.2 Selecting Views Using MOSA

In our representation for view selection problem using MOSA, a solution S_0 is said to dominate a solution S_1 if total query processing cost using S_0 and total view maintenance cost for solution S_0 for an MVPP is less than or equal to respective total costs for S_1 and at least one of these costs for S_0 is less than that for S_1.

At every iteration of every temperature epoch, from a current solution S_0, by perturbing, a new solution S_1 is generated, whose total space requirement is less than or

equal to the maximum space(constraint) specified, as represented in sub-section 4.1. The dominance between these two solutions are checked as defined in (12) in section 3. At a particular iteration, of particular temperature level t , in this algorithm, let f_0 be the number of solutions in the archive that dominates S_0, f_1 be the total number of solutions in the archive that dominates the solution S_1 and f_a be the number of solutions in the archive and current solution. Now, if $S_0 \prec S_1$, then to set $S_0 = S_1$ with

$$\text{probability} = \min(1, \exp\{-((f_1 - f_0)/f_a)/t\}) \tag{13}$$

If $S_0 \nprec S_1$ and $S_1 \nprec S_0$, but k ($k \geq 1$) number of solutions in the archive dominate S_1 then to set $S_0 = S_1$ with probability given by (13). But if $S_0 \nprec S_1$ and $S_1 \nprec S_0$ and also S_1 does not dominate any solution in the archive so far, then to set $S_0 = S_1$ and append S_1 to the archive. Again if $S_0 \nprec S_1$ and $S_1 \nprec S_0$ and S_1 dominates k ($k \geq 1$) number of solutions in the archive then to set $S_0 = S_1$ and append S_1 to the archive and to remove all k number of solutions dominated by S_1.

In the case of $S_1 \prec S_0$ and k ($k \geq 1$) number of solutions in the archive dominates S_1, we have to set $S_0 = S_1$ with probability given by (13). But if $S_1 \prec S_0$ and also S_1 dominates k ($k \geq 1$) number of solutions in the archive, set $S_0 = S_1$ and append S_1 to the archive and to remove all k number of solutions dominated by S_1. If $S_1 \prec S_0$ but S_1 does not dominate any solution in the archive, then to set $S_0 = S_1$ and append S_1 to the archive and if S_0 is already in the archive then it is to be removed from the archive.

The above process continues for a number of iterations or till it reaches the terminating condition specified, and at the end of this loop, the temperature is decreased by a specified amount. If the temperature is still above the terminating temperature, then again for the specified number of iterations or terminating condition, the process continues. Thus at the end, a set of non-dominating solutions of the problem will be yielded.

4.3 Selecting Views Using AMOSA

To implement AMOSA, as suggested by Sanghamitra Bandyopadhyay et al. in [29], an archive to keep the non dominating solutions is initialized at the beginning. Then by using the array containing the labels of temporary views of the MVPP and their respective sizes, a random set of views are selected, whose total space requirement for materializing is less than or equal to the maximum space(constraint) specified. This set, which is represented as a string of bits as discussed in sub-section 4.1, is considered as current solution, S_0 so far, and added to the archive. Then the temperature is reduced from initial maximum temperature, T_{max}, by the rate specified, α, and in each temperature the following steps are executed for a fixed number of iterations till it does not reach the minimum terminating temperature, T_{min}. Thus, for a number of iterations, the AMOSA works as follows:

- The current solution S_0 is perturbed to generate a new candidate solution, S_1 satisfying the space constraint.
- Then by computing, $Q_G(S_0)$, $Q_G(S_1)$, $U_G(S_0)$ and $U_G(S_1)$ the domination between S_0 and S_1 is checked as defined by (12).

Now based on the domination type,

- If $S_0 \prec S_1$ and some solution of the archive dominates S_1, then set $S_0=S_1$, for next iteration with probability,

$$\frac{1}{1+\exp(((\sum_{i=1}^{k} \Delta dom_{i,new\_sol}) + \Delta dom_{current\_sol,new\_sol}) /(k+1)) * temp)}$$

- If $S_0 \not\prec S_1$ and $S_1 \not\prec S_0$, then
  - If S_1 is dominated by $k(k\geq 1)$ points of the archive then set $S_0=S_1$, for next iteration with probability

$$\frac{1}{1+\exp(((\sum_{i=1}^{k} \Delta dom_{i,new\_sol}) / k) * temp)}$$

  - If new solution S_1 is non-dominating with any solution in the archive, then set $S_0=S_1$ and S_1 is to be added to the archive.
  - But if S_1 dominates k ($k\geq 1$) solutions of the archive then we set $S_0=S_1$ and S_1 is added to the archive and all k points in the archive dominated by S_1 are removed.
- If $S_1 \prec S_0$, then
  - If k ($k\geq 1$) solutions of the archive dominates the new solution S_1, then the minimum of the *difference of domination* amounts between S_1 and the k solutions of the archive that dominate S_1, Δdom_{min}, is calculated and the solution which correspond to this Δdom_{min} is set as S_0 with probability equals to $1/(1+\exp(-\Delta dom_{min}))$.
  - If *new solution* S_1 is non-dominating with respect to the points in the archive, except the current solution S_0, if it is in the archive, the new solution S_1 is appended to the archive and the current solution S_0 is to be removed from the archive, if it is present in the archive. Then set $S_0=S_1$.
  - If the new solution S_1 also dominates k ($k\geq 1$) solutions already present in the archive, then set $S_0=S_1$ and S_1 is added to the archive and all k points of archive dominated by the new solution S_1 are removed.
- The process continues for a number of iterations for each temperature (*temp*) and the temperature is reduced in a cooling rate say α till it reaches a predefined minimum temperature, say T_{min}.

Thus finally an archive of non-dominating solution to the problem is achieved. As the size of the archive is limited, therefore, if the size of the archive crosses the limit, clustering technique may be used to reduce the number of solution in the archive.

5 Experiments and Analysis

To analyze the effectiveness of multi-objective simulated annealing in view selection for materializing in data warehouses, we have implemented MOSA and AMOSA for TPC-H benchmark[30] data warehouse. The TPC-H schema was built and loaded

with data using "dbgen" utility of TPC-H framework in Oracle10g RDBMS. The TPC-H bench mark queries were generated using "qgen" utility of TPC-H framework. We selected those queries of TPC-H which go well with our application. Similarly some base relations are used as original and some are slightly changed to suit our experiment. The comparative results of MOSA and AMOSA based technique to deal with this problem in terms of quality of solutions are presented in figures Fig. 1 and Fig. 2 by using a simple MVPP of three queries with sixteen temporary intermediate views from ❄❄ benchmark data warehouse. We used "explain-plan" utility of Oracle RDBMS for finding costs of intermediate views. For computing gross query processing cost for an MVPP and view maintenance cost of the MVPP, we used the algorithm suggested by Yang, et al. in [17]. In our experiment it is found that the quality of solutions in AMOSA is better than MOSA as in data warehouse scenario, frequency of view maintenace is negligible compared to query frequency.

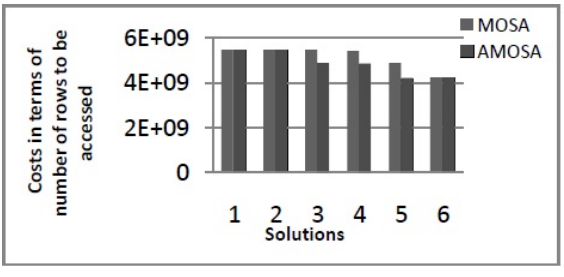


Fig. 1. Query processing costs for a set of views selected as solutions

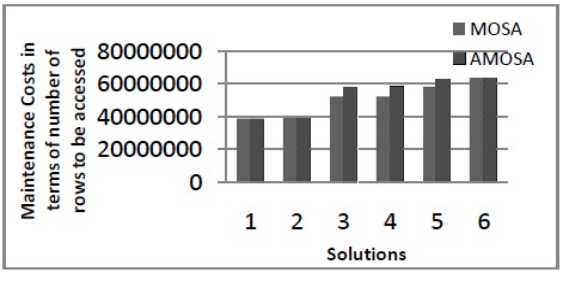


Fig. 2. View maintenance costs for a set of views selected as solutions

6 Conclusions

As mostly the view selection for materialization problem is dealt as single objective optimization problem, here we handled the problem as multi-objective optimization problem. In this paper we have presented multi-objective simulated annealing based technique for view selection problem. We tried to deal the problem with a recent multi-objective simulated annealing technique termed as AMOSA and found that it works better than MOSA. Though so far we tried with very small and simple data set from TPC-H, it may be well tested with very large and complex MVPP.

As a future work we intend to do the experimentation with very large and complex MVPP containing very large number of temporary views by using other randomized and stochastic method to achieve further improvement in the quality of solutions.

References

1. Gupta, A., Mumick, I.S.: Maintenance of Materialized Views: Problems, Techniques, and Applications. ACM (1999) ISBN:0-262-57122-6
2. Vijay Kumar, T.V., Aloke, G.: Greedy Selection of Materialized Views. Int. J. of Computer and Communication Technology 1(1), 47–58 (2009)
3. Harinarayan, V., Rajaraman, A., Ullman, J.: Implementing Data Cubes Efficiently. In: Proceedings of ACM SIGMOD International Conference on Management of Data (1996)
4. Gupta, H., Harinarayan, V., Rajaraman, A., Ullman, J.D.: Index Selection for OLAP. In: 13th ICDE Conference, pp. 208–219 (1997)
5. Derakhshan, R., Dehne, F., Korn, O., Stantic, B.: Simulated Annealing for Materialized View Selection in Data Warehousing Environment. In: Proceedings of the 24th IASTED International Conference on Database and Applications, pp. 89–94 (2006)
6. Derakhshan, R., Stantic, B., Korn, O., Dehne, F.: Parallel Simulated Annealing for Materialized View Selection in Data Warehousing Environments. In: Bourgeois, A.G., Zheng, S.Q. (eds.) ICA3PP 2008. LNCS, vol. 5022, pp. 121–132. Springer, Heidelberg (2008)
7. Gupta, H., Mumick, I.S.: Selection of Views to Materialize under a Maintenance Cost Constraint. In: Beeri, C., Bruneman, P. (eds.) ICDT 1999. LNCS, vol. 1540, pp. 453–470. Springer, Heidelberg (1998)
8. Gupta, H.: Selection and maintenance of views in a data warehouse. PhD thesis, Stanford University (1999)
9. Nadeua, T.P., Teorey, T.J.: Achieving Scalability in OLAP Materialized View Selection. In: DOLAP 2002, pp. 28–34. ACM (2002)
10. Agrawal, S., Chaudhuri, S., Narasayya, V.: Automated Selection of Materialized Views and Indexes for SQL Databases. In: 26th VLDB Conference, Cairo, Egypt (2000)
11. Chan, G.K.Y., Li, Q., Feng, L.: Optimized Design of Materialized Views in a Real-Life Data Warehousing Environment. International Journal of Information Technology 7(1), 30–54 (2001)
12. Horng, J.T., Chang, Y.J., Liu, B.J., Kao, C.Y.: Materialized View Selection Using Genetic Algorithms in a Data Warehouse System. In: IEEE CEC (1999)
13. Labrinidis, A., Roussopoulos, N.: Web View Materialization. In: ACM SIGMOD Conference, Dallas, Texas, USA, pp. 367–378 (2000)
14. Loureiro, J., Belo, O.: An Evolutionary Approach to the Selection and Allocation of Distributed Cubes. In: IEEE IDEAS 2006, pp. 243–248 (2006)
15. Saidi, S., Slimani, Y., Arour, K.: Web View Selection from User Access Patterns. In: PIKM 2007, Lisboa, Portugal, pp. 171–176 (2007)
16. Serna-Encinas, M.T., Hoya-Montano, J.A.: Algorithm for selection of materialized views: based on a costs model. In: Proceeding of Eighth International Conference on Current Trends in Computer Science, pp. 18–24 (2007)
17. Yang, J., Karlapalem, K., Li, Q.: Algorithm for Materialized View Design in Data Warehousing Environment. In: VLDB 1997, pp. 136–145 (1997)
18. Zhang, C., Yang, J.: Genetic Algorithm for Materialized View Selection in Data Warehouse Environments. In: Mohania, M., Tjoa, A.M. (eds.) DaWaK 1999. LNCS, vol. 1676, pp. 116–125. Springer, Heidelberg (1999)

19. Zhang, C., Yao, X., Yang, J.: An Evolutionary Approach to Materialized Views Selection in a Data Warehouse Environment. IEEE Transactions on Systems and Cybernetics Part C: Applications and Reviews 31(3), 282–294 (2001)
20. Lee, M., Hammer, J.: Speeding up Materialized View Selection in Data Warehouses Using a Randomized Algorithm. Int. J. Cooperative Inform. Syst. 10, 327–353 (2001)
21. Sun, X., Wang, Z.: An Efficient Materialized Views Selection Algorithm Based on PSO. In: Proc. International Workshop on Intelligent Systems and Applications (2009) Print ISBN: 978-1-4244-3893-8
22. Zhang, Q., Sun, X., Wang, Z.: An Efficient MA-Based Materialized Views Selection Algorithm. In: Proceedings of the 2009 IITA International Conference on Control, Automation and Systems Engineering, CASE 2009 (2009) ISBN: 978-0-7695-3728-3
23. Aouiche, K., Jouve, P.-E., Darmont, J.: Clustering-Based Materialized View Selection in Data Warehouses. In: Manolopoulos, Y., Pokorný, J., Sellis, T.K. (eds.) ADBIS 2006. LNCS, vol. 4152, pp. 81–95. Springer, Heidelberg (2006)
24. Aouiche, K., Darmont, J.: Data mining-based materialized view and index selection in data warehouses. Journal of Intelligent Information System 33, 65–93 (2009)
25. Das, A., Bhattacharyya, D.K.: Density-Based View Materialization. In: Pal, S.K., Bandyopadhyay, S., Biswas, S. (eds.) PReMI 2005. LNCS, vol. 3776, pp. 589–594. Springer, Heidelberg (2005)
26. Davoud, S., Ellips, M.: Particle Swarm Optimization Methods, Taxonomy and Applications. International Journal of Computer Theory and Engineering 1(5), 486–502 (2009)
27. Smith, K.I., Everson, R.M., Fieldsend, J.E., Murphy, C., Misra, R.: Dominance-Based Multiobjective Simulated Annealing. IEEE Transactions on Evolutionary Computation 12(3), 323–342 (2008)
28. Smith, K., Everson, R., Fieldsend, J.: Dominance measures for multi-objective simulated annealing. In: Proc. 2004 IEEE Congr. Evol. Comput., pp. 23–30 (2004)
29. Bandyopadhyay, S., Saha, S., Maulik, U., Deb, K.: A Simulated Annealing Based Multiobjective Optimization Algorithm: AMOSA. IEEE Transactions on Evolutionary Computation 12(3), 269–283 (2008)
30. Transaction Processing Performance Council: in TPC Benchmark^{TM} (Decision Support), Standard Specification Revision 2.14.2, `http://www.tpc.org/tpch/` (accessed July 20, 2011)

Comparison of Deterministic and Probabilistic Approaches for Solving 0/1 Knapsack Problem

Ritika Mahajan[1], Sarvesh Chopra[2], and Sonika Jindal[3]

[1] M.Tech (Computer Science Engineering), Shaheed Bhagat Singh College of Engineering and Technology, Ferozepur, India
er.ritikamahajan@gmail.com
[2] M.Tech (Information Technology), Guru Nanak Dev Engineering College, Ludhiana, India
er.sarveshchopra@gmail.com
[3] Assistant Professor, Department of Computer Science and Engineering, Shaheed Bhagat Singh College of Engineering and Technology, Ferozepur, India
sonikamanoj@gmail.com

Abstract. The purpose of this paper is to analyze algorithm design paradigms applied to single problem – 0/1 Knapsack Problem. The Knapsack Problem is a combinatorial optimization problem where one has to maximize the benefits of objects in a knapsack without exceeding its capacity. It is an NP-complete problem and uses exact and heuristic techniques to get solved.

The objective is to analyze that how the various techniques like Dynamic Programming and Genetic Algorithm affect the performance of Knapsack Problem. Our experimental results show that the promising approach is genetic algorithm as it gives result in optimal time.

Keywords: Knapsack Problem, NP-complete problem, Dynamic Programming, Genetic Algorithm.

1 Introduction

The Knapsack problem is a combinatorial optimization problem where one has to maximize the benefit of objects in a knapsack without exceeding its capacity. Given a set of items we have to find optimal packing of a knapsack. Each item is characterized by weight and value and knapsack is characterized by capacity. Optimal packing is the one in which weight is less or equal to the capacity and in which value is maximal among other feasible packings. [1]

We have n kinds of items, 1 through n. Each kind of item i has a value v_i and a weight w_i. All values and weights are nonnegative. The maximum weight that we can carry in the bag is W.

More formally:

given a number of items n, their weights $W = w_1 \ldots\ldots w_n$,
their values $V = v_1 \ldots\ldots v_n$ and knapsack capacity c,
find vector $X = x_1 \ldots.. x_n$ so that $(x_1 * w_1 + \ldots.. x_n * w_n) <= c$
and $(x_1 * w_1 + \ldots.. x_n * w_n)$ is maximal. [2]

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 629–637.
springerlink.com

There are numerous versions to this problem. We shall consider only three: [3]

- 0/1 Problem
- Bounded Problem
- Unbounded Problem

0-1 knapsack problem, which restricts the number x_i of copies of each kind of item to zero or one.

Mathematically the 0-1-knapsack problem can be formulated as:

- maximize $\sum_{i=1}^{n} v_i x_i$
- subject to $\sum_{i=1}^{n} w_i x_i \leqslant W, \qquad x_i \in \{0,1\}$

The **bounded knapsack problem** restricts the number x_i of copies of each kind of item to a maximum integer value c_i. Mathematically the bounded knapsack problem can be formulated as:

- maximize $\sum_{i=1}^{n} v_i x_i$
- subject to $\sum_{i=1}^{n} w_i x_i \leqslant W, \qquad x_i \in \{0,1,\ldots,c_i\}$

The **unbounded knapsack problem** (**UKP**) places no upper bound on the number of copies of each kind of item, i.e. $x_i = 0\ldots$infinity.

1.1 Existing Techniques for Knapsack Problem

Two types of algorithms exist for NP hard problems:

1. Exact Algorithms: Algorithms that find exact solutions (they will work reasonably fast only for relatively small problem sizes).
2. Heuristic Algorithms: Algorithms that deliver either seemingly or probably good solutions but which could not be proved to be optimal.

2 Metric Designs

Metrics are the set of measurements that quantify results. Quantitative measurement helps to estimate algorithm quality and complexity. Metric give us objective information about the properties of algorithm. Algorithm Metrics are the measures that can be used to determine the quality of an algorithm. Here metrics are quantitative analysis of the usage on the basis of different parameters.

This metric set is formed with the help of report generated by running different programs:

- **Space Complexity:** It is representing the memory size in bytes that the algorithm is taking to run. When the algorithm executes, different algorithms require different memory requirement.
- **Time Complexity:** It is representing the time in seconds the algorithm is taking to run. It tells how much time is the algorithm taking to execute and to give the optimal solution.
- **Number of operations (NOP):** It tells that how many numbers of operations the algorithm will perform. It include all the executable statements, loops, decision making statements etc.
- **Best Value:** It tells that which algorithm is giving best value that is maximum value or maximum profit.
- **Programming Effort :** When the algorithm is converted into program to run by using some programming language like c++, then the algorithm which require less programming effort is considered as best.
- **Degree of nearness to optimality:** As we cannot get the exact optimal solution for NP complete problem, so we are considering the algorithm which is giving near to optimal solution as a best.
- **Degree of user friendliness for implementation:** It is user friendliness of program that which program is easier to understand and to implement.

The best algorithm is decided on the basis that which takes less time to run and gives maximum profit.

3 Dynamic Programming

Dynamic programming (DP), the most powerful design technique for optimization problems was invented by Richard Bellman, a prominent U.S. mathematician, in 1950s. The solutions for the dynamic programming are based on multistage optimizing decisions on a few common elements**.** Here programming does not symbolize computer programming but "planning". The DP is closely related to divide and conquer technique where the problem breaks down into smaller subproblems and each subproblem is solved recursively. The DP differs from divide and conquer in a way that instead of solving subproblem recursively, it solve each of the subproblems only once and store the solution to the subproblems in a table. Later on, the solution to the main problem is obtained by these subproblems solutions. [1][4]

3.1 Algorithm

(Weights [1 … N], Values [1 … N],Table [0 ... N, 0 … Capacity])[5]
// Input:
Array Weights contains the weights of all items
Array Values contains the values of all items
Array Table is initialized with 0s; it is used to store the results from the dynamic programming algorithm.

// Output:

The last value of array Table (Table [N, Capacity]) contains the optimal solution of the problem for the given Capacity.

```
for i = 0 to N do
for j = 0 to Capacity
        if j < Weights [i] then
                Table [i, j] ← Table[i-1, j]
        else
                Table [i, j] ← maximum { Table[i-1, j]
                                AND
                        Values [i] + Table [i-1, j – Weights [i]]
return Table[N, Capacity]
```

In the implementation of the algorithm instead of using two separate arrays for the weights and the values of the items, we used one array Items of type item, where item is a structure with two fields: weight and value.

To find which items are included in the optimal solution, we use the following algorithm:

```
n ← N
c ← Capacity
Start at position Table [n, c]
While the remaining capacity is greater than 0 do
If Table [n, c] = Table [n-1, c] then
Item n has not been included in the optimal solution
Else
Item n has been included in the optimal solution
Process Item n
Move one row up to n-1
Move to column c – weight (n)
```

4 Genetic Algorithm

A genetic algorithm is a computer algorithm that searches for good solutions to a problem from among a large number of possible solutions. All Gas begin with a set of solutions (represented by chromosomes) called population. A new population is created from solutions of an old population in hope of getting a better population. Solutions which are then chosen to form new solutions (off springs) are selected according to their fitness. The more suitable the solutions are the bigger chances they have to reproduce. This process is repeated until some condition is satisfied. Most Gas methods are based on the following elements: “populations of chromosomes, selection according to fitness, crossover to produce new offspring, and random mutation of new offspring”. [6][7].

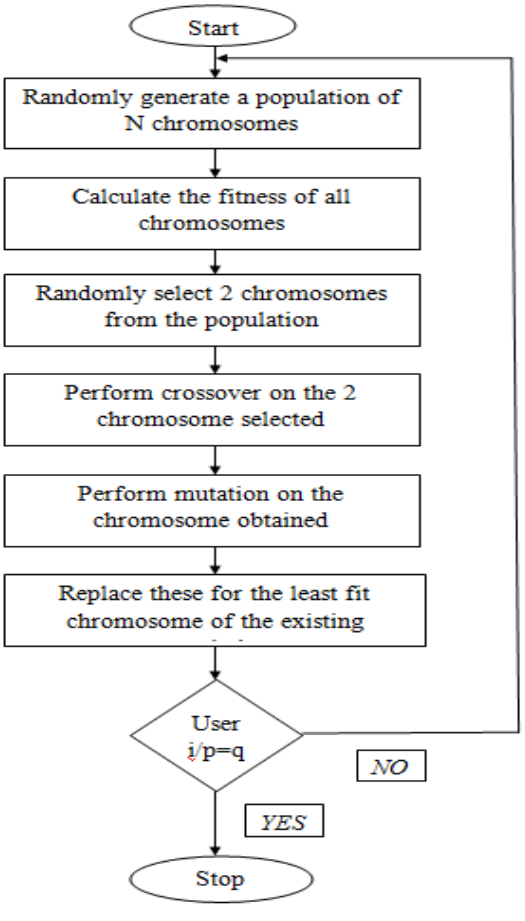
These algorithms can use for finding a suitable solution, but these algorithms do not necessarily provide the best solution. The solutions found by these methods are often considered as good solutions, because it is not often possible to prove what the optimum is. [6]

A genetic algorithm (GA) is a search technique used in computer science to find approximate solutions to optimization and search problems. Genetic algorithms are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, natural selection, and recombination (or crossover). [6]

4.1 Algorithm

1. Randomly generate population of N chromosomes.
2. Calculate the fitness of each chromosome.
3. Randomly select two chromosomes from the population.
4. Perform the crossover on two chromosomes selected from the population.
5. Perform the mutation on the chromosomes obtained.
6. Replace these for the least fit chromosomes in the exiting population.
7. Repeat steps 2 to 6 until optimal solution is found.

4.2 Flowchart of Genetic Algorithm



5 Results and Comparisons

For the testing of the different algorithms, files are generated with different sizes where each record consists of a pair of randomly generated integers representing the weight and value of each item. We performed the testing in which number of items is increased, while the capacity of the knapsack constant (=50).

Table 1. For genetic algorithm, items are increasing but capacity remains constant (=50)

Items	Weight	Value	Time	Memory	No. of operations
5	30	75	0.7692	1676	719
10	50	103	1.3626	1676	1899
15	48	80	38.2198	1676	24755
20	50	60	53.4945	1676	22119
25	50	59	71.8681	1676	24225

Table 2. For dynamic approach, items are increasing but capacity remains constant (=50)

Items	Weight	Value	Time	Memory	No. of operations
5	20	50	41.3736	1472	316
10	22	44	64.3406	1472	581
15	49	62	71.5385	1472	846
20	49	56	88.3516	1472	1111
25	50	54	184.560	1472	1376

Table 3. Comparison, when items = 5 and capacity is constant (=50)

Approach	Value	Time	Memory	No. of operations
Genetic Algorithm	75	0.7692	1676	719
Dynamic Approach	50	41.3736	1472	316

Table 4. Comparison, when items = 10 and capacity is constant (=50)

Approach	Value	Time	Memory	No. of operations
Genetic Algorithm	103	1.3626	1676	1899
Dynamic Approach	44	64.3406	1472	581

Table 5. Comparison, when items = 15 and capacity is constant (=50)

Approach	Value	Time	Memory	No. of operations
Genetic Algorithm	80	38.2198	1676	24755
Dynamic Approach	62	71.5385	1472	846

Table 6. Comparison, when items = 20 and capacity is constant (=50)

Approach	Value	Time	Memory	No. of operations
Genetic Algorithm	60	53.4945	1676	22119
Dynamic Approach	56	88.3516	1472	1111

Table 7. Comparison, when items = 25 and capacity is constant (=50)

Approach	Value	Time	Memory	No. of operations
Genetic Algorithm	59	71.8681	1676	24225
Dynamic Approach	54	184.560	1472	1376

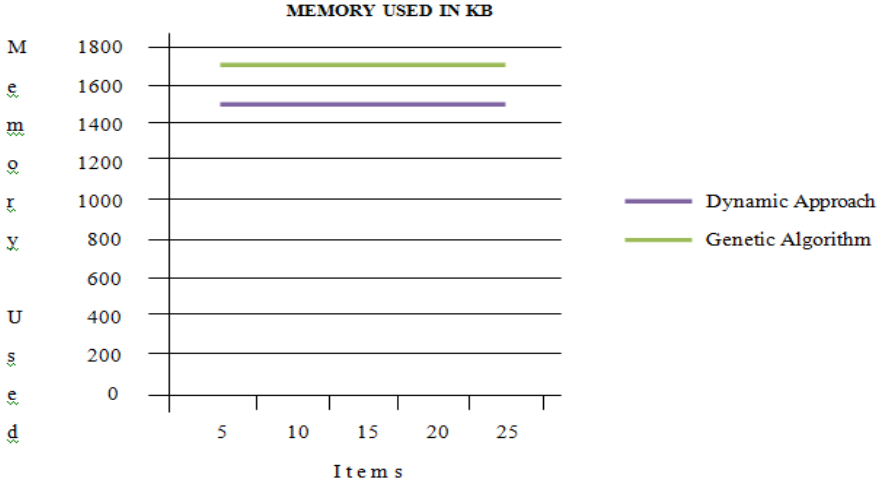


Fig. 1. This figure shows that as the items vary, the memory requirement (in KB) remains constant. Genetic Algorithm is demanding large memory but dynamic programming requires very less memory.

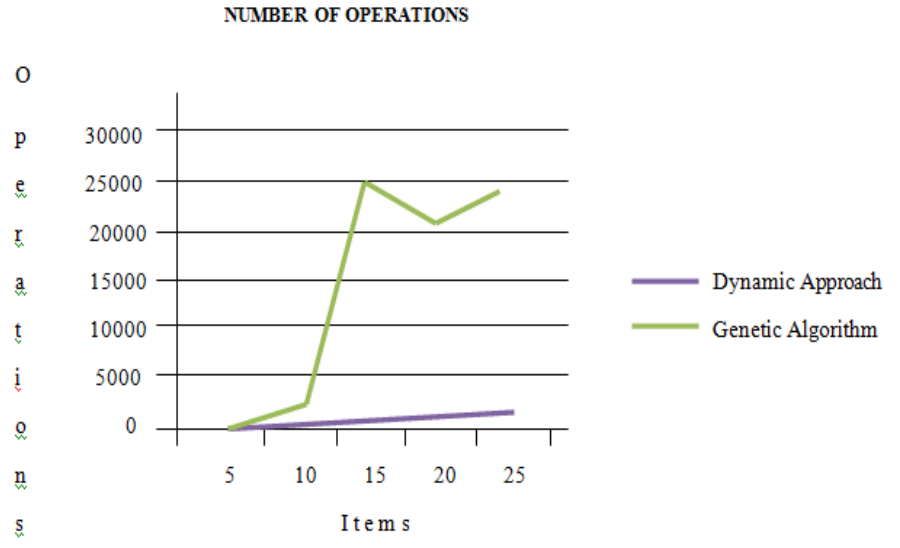


Fig. 2. This figure shows that Genetic Algorithm perform very large number of basic operations. The number of basic operations performed by Dynamic programming is negligible as compared with Genetic Algorithm.

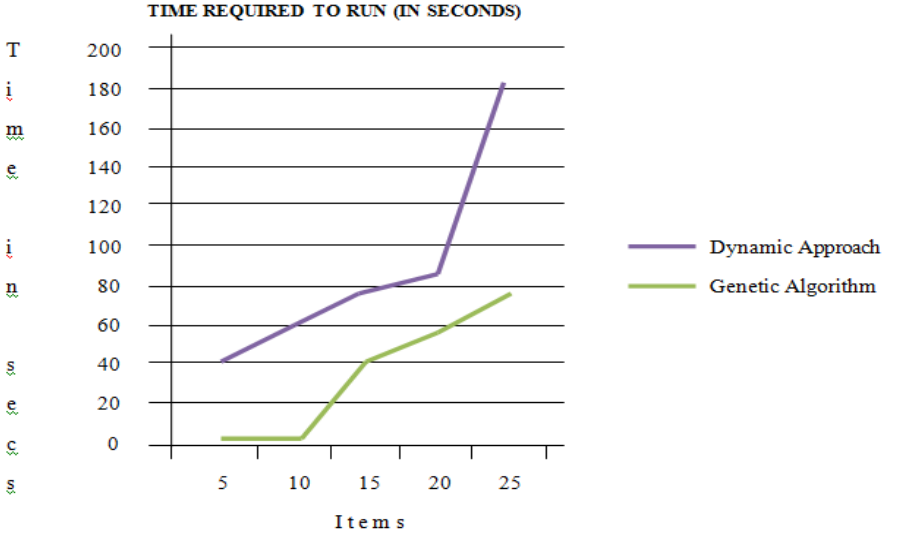


Fig. 3. This figure shows that Genetic Algorithm executes in very less time (in seconds). Dynamic Programming takes more time. When the number of items increases, their execution time also increases. Moreover that algorithm is considered as best which takes less time to execute.

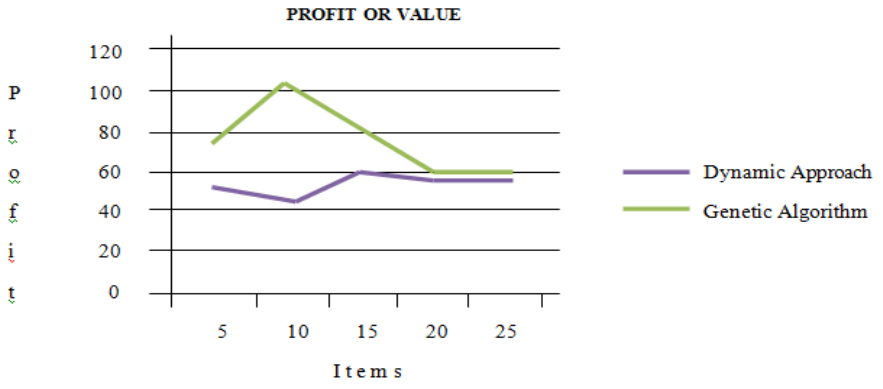


Fig. 4. This figure shows that Genetic Algorithm is giving more profit to knapsack In knapsack Problem, when the items are put in knapsack, the major goal is that all the items which are in knapsack should be of high profit. As genetic algorithm is making highest profit so it is considered as best.

6 Conclusion

We can conclude that Genetic Algorithm is giving more profit to knapsack problem as compared with dynamic approach. Genetic Algorithm, in spite of taking large amount of memory, takes very less time to execute. The only major drawback of Genetic Algorithm is that it requires large programming efforts.

So Genetic Algorithm can be considered as best algorithm for finding the near to optimal solution for the most widely used NP-Complete problem i.e. Knapsack Problem if there is no constraint of memory. For getting the near to optimal solution in less time, we can use the Genetic Algorithm.

Acknowledgment. We express our sincere gratitude to Mrs. Sonika Jindal who helped us immensely in completing this paper with her guidance.

References

Cormen, T.H., Leiserson, C.E., Riverst, R.L., Stein, C.: Introduction to Algorithms

Knapsack problem, http://www.utdallas.edu/~scniu/OPRE-6201/documents/DP3-Knapsack.pdf

Poirriez, V., Yanev, N., Andonov, R.: A Hybrid Algorithm for the Unbounded Knapsack Problem (October 28, 2008)

Dynamic programming, http://en.wikipedia.org/wiki/Dynamic_programming

0/1 knapsack problem, http://www.brpreiss.com/books/opus4/html/page445.html

Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley (1990)

Mitchell, M.: An Introduction to Genetic Algorithms. MIT Press

Thesis by Shailendra Kumar, Choosing Best Algorithm Design Strategies For a Particular Problem (June 2009)

Comparison of Content Based Image Retrieval System Using Wavelet Transform

Suchismita Das, Shruti Garg, and G. Sahoo

Department of Computer Science and Engineering
Birla Institute of Technology, Mesra, Ranchi
sd.suchi@gmail.com, gshruti_garg@yahoo.com,
gsahoo@bitmesra.ac.in

Abstract. The large numbers of images has posed increasing challenges to computer systems to store and manage data effectively and efficiently. This paper implements a CBIR system using different feature of images through four different methods, two were based on analysis of color feature and other two were based on analysis of combined color and texture feature using wavelet coefficients of an image. To extract color feature from an image, one of the standard ways i.e. color histogram was used in YCbCr color space and HSV color space. Daubechies' wavelet transformation and Symtels' wavelet transform were performed to extract the texture feature of an image. After obtaining all experimental results, it has been inferred that wavelet based method gave a better performance as compared to color based method.

Keywords: CBIR, wavelet transformation, Color histogram, YCbCr, HSV.

1 Introduction

Now days, CBIR (content based image retrieval) is a hotspot of digital image processing techniques. CBIR research started in the early 1990's and is likely to continue during the first two decades of the 21st century [11]. There is a growing interest in CBIR because of the limitations inherent in metadata-based systems, as well as the large range of possible uses for efficient image retrieval. The term 'content' in this context might refer to colors, shapes, textures, or any other information that can be derived from the image itself. Content based image retrieval (CBIR) is therefore proposed, which finds images that have visual low-level image features similar to those of the query image example [8].

There are two approaches to image retrieval: Text-Based approach and Content-Based approach. Text-Based approach has some obvious shortcomings as each person can have different perception for each textual description. It is also time consuming when dealing with very large databases. Content based retrieval of visual data requires a paradigm that differs significantly from both traditional databases and text based image understanding systems [8]. "Content-based" means that the search will analyze the actual contents of the image rather than the metadata such as keywords, tags, and/or descriptions associated with the image.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 639–648.
springerlink.com

Feature extraction is very crucial step in image retrieval system to describe the image with minimum number of descriptors. The basic visual features of images include color and texture [9]. The color gives user a feel in terms of visual similarity but the texture does not give much of visual feel but it helps to retrieve [6] based on patterns / textures [1, 5]. The histogram is the most commonly used structure to extract the color component of an image [8]. Texture is another important property of images. Various texture representations have been investigated in pattern recognition and computer vision. Wavelet transformations were the most useful technique to extract the texture feature of an image.

In CBIR each image that is stored in the database has its features extracted and compared to the features of the query image [7]. CBIR systems can also make use of relevance feedback, where the user can progressively refines the search results by marking images in the results.

2 Proposed Schemes

In this paper four methods of image retrieval were proposed, out of which two were based on color feature of an image and other two were based on combined color and texture feature of an image.

2.1 Image Retrieval System Based on Color Feature

Content based image retrieval system based on color similarity is achieved by computing a color histogram for each image that identifies the proportion of pixels having specific values within an image (that humans express as colors). The proposed color based image retrieval technique is a certain modification of the methods referred in [11].

2.1.1 System Using HSV Color Model

HSV color space is widely used in computer graphics, visualization in scientific computing and other fields [3]. The color-space of the image is changed from RGB to a space known as Hue Saturation Value (HSV). A three dimensional representation of the HSV color space is a hexacone, where the central vertical axis represents the Intensity [2]. In the HSV color space each pixel contributes its hue and intensity based on its saturation. The generated histogram consists of “true color” components and “gray color” components, which store contributions from the hue and the intensity of each pixel. As hue varies from 0 to 360 degree, the corresponding colors vary from red through yellow, green, cyan, blue, magenta, and back to red. As saturation varies from 0 to 1.0, the corresponding colors (hues) vary from unsaturated (shades of gray) to fully saturated (no white component). As value, or brightness, varies from 0 to 1.0, the corresponding colors become increasingly brighter.

2.1.2 System Using YCbCr Color Model

The more frequently adopted approach for CBIR systems is based on the conventional color histogram (CCH),which contains occurrences of each color in a particular color space obtained counting all image pixels having that color. The most common color representation model is RGB color model in which colors are represented as a combination of various intensities of red, green, and blue. Another color space, widely used

for digital image and video is the YCbCr color space (also known as YUV).In this color space luminance (brightness or intensity) information is stored as a single component (Y). Chrominance (color) information is stored as two color-difference components (Cb and Cr). Cb represents the difference between the blue component and a reference value. Cr represents the difference between the red component and a reference value.

Comparing all the colors between two images would be very time consuming and complex. So reducing the amount of information is performed by quantizing the color distribution into color histograms. When computing a color histogram for an image, the different color axes are divided into a number of so-called bins [7].

2.1.3 Algorithm for Color Based Image Retrieval

The steps involved in color based image retrieval are given below:

Step1: Preprocessing: All images in database were resized to a fixed size of 160×120 pixels to make all images similar size.

Step2: Each image of the database was then converted to HSV/YCbCr color space.

Step3: Color histogram (CCH) was computed for each of the three planes and Quantization was performed. Color histogram was computed by joint probabilities of intensities of the color channels. The color histogram can be thought of as a set of vectors.

More formally, the color histogram is defined by,

$$h_{A,B,C}(a,b,c) = N \cdot Prob(A = a, B = b, C = c)$$

Where A, B and C represent the three color channels (H, S, V or YCbCr) and N is the number of pixels in the image [2].

Step 4: A feature vector was formed based on these values of standard Deviation which was used during image retrieval process for similarity measure.

The standard deviation value was calculated as:

$$S = \sqrt[2]{\left(\frac{1}{m}\sum(x_i - \bar{x})^2\right)} \quad (1)$$

Where, $x = 1/m \sum_{i=1}^{m} x_i$

Where, m=No. of elements in the sample.

2.2 Image Retrieval System Based on Color and Texture Feature

A very basic issue in designing a CBIR system is to select the most effective image features to represent image contents [10]. In present paper content based image retrieval system was designed by combining both color and texture. Here texture was represented using wavelet coefficients. This was a certain alteration of the method proposed in [4] for texture feature and in [11] for color feature extraction.

Mathematically, wavelet transform is a convolution operation, which can equivalent to passing the pixel values of an image through a low pass filter and a high pass

filter. Although suggested by some researchers and are easier to implement, Haar wavelets do not have sufficiently sharp transition and hence are not able to separate different frequency bands appropriately. Daubechies' wavelets, on the other hand, have better frequency resolution properties because of their longer filter lengths [5].

2.2.1 Texture Feature Extraction

In this paper two wavelet methods i.e. Daubechies and Symtel, were implemented for image retrieval where three wavelet coefficients LH, HL and HH were used. The high level frequency components contain information about edges and high level imaged details, that's why all high level coefficients were taken for matching.

Each image was decomposed into four subbands and 5 level wavelet transform was performed, and only the detailed coefficient of each level were taken. The approximation coefficient of the image had been ignored.

2.2.2 Algorithm for Proposed Scheme

Content based image retrieval system using above two methods i.e. using symtel's wavelet transform and, Daubechies' Wavelet Transform were implemented through following steps only with the difference that each method used different wavelet transform method as mentioned above.

Step1: Preprocessing: All images in database were resized to a fixed size of 160×120 pixels to make all images similar size.

Step2: Color feature was extracted by computing color histogram for each of the color plane (red, green, blue) of an image in RGB color space.

Step 3: Standard deviation of each histogram corresponding to each color plane was computed. Hence each image had three elements for color feature in feature vector.

Step4: To extract texture feature, Daubechies/Symtel wavelet transformation was performed to 5 levels and the coefficient of each level was taken as described in methods above.

Step5: After getting coefficients, Standard deviation of each coefficient is computed for each image of the database.

Step6: A feature vector was formed with 18 elements out of which 3 elements were based on color feature and 15 elements were based on texture feature. The feature vector was then stored in the database for image retrieval purposes as similarity measure.

3 Similarity Measure

Content-based image retrieval determines visual similarities between a query image and images in a database. Different similarity measures will affect retrieval performances of an image retrieval system significantly. In this paper three different distance metric to measures were used and the performance evaluated by each method.

Canberra and Euclidean distance, both measures were used for finding similar images. Euclidean distance is used for all three methods between the standard deviations of the query image and the images in the database.

$$\text{Euclidean distance} = \sqrt{\sum(a_i - b_i)^2} \tag{2}$$

$$\text{Canberra distance} = \sum |a_i - b_i|/|a_i| + |b_i| \tag{3}$$

4 Experimental Results

Around 1000 images were stored in the database, which consists of 12 different categories namely red rose, bird, sunflower, girl, aquarium, car, girl in rain, starfish, baby, tiger, horse and oranges. Out of these 12 categories of images, the results of tiger and bird are shown below. In each of the results shown below the top middle of the window shows the query image.

Fig.1.a HSV Color Based Method

Fig. 1.b YCbCr Color Based Method

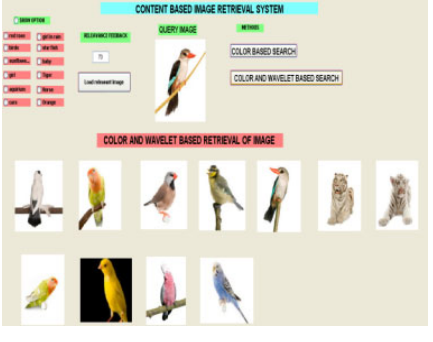


Fig. 1.c Symtel Wavelet Transform Method

Fig. 1.d Daubechies Wavelet Transform Method

Fig. 2.a HSV Color Based Method

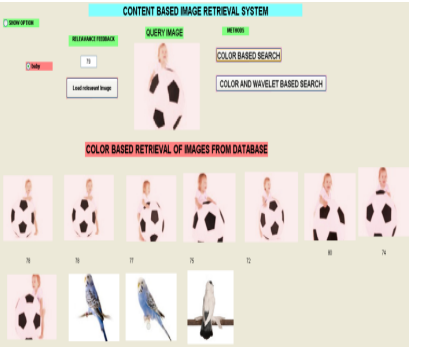


Fig. 2.b YCbCr Color Based Method

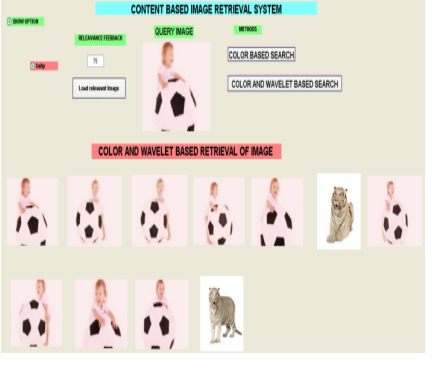


Fig. 2.c Symtel Wavelet Transform Method

Fig. 2.d Daubechies Wavelet Transform Method

Performance comparison of content-based image retrieval systems is a crucial and non-trivial task since it is very difficult to determine the relevant sets. The commonly used performance measurement parameters for the evaluation of retrieval performance are, precision and recall [11]. Precision *P* measures the accuracy of the retrieval. Precision *P* is defined as the ratio of the number of retrieved relevant images to the total number of retrieved images. Recall R measures the robustness of the retrieval. Recall R is defined as the ratio of the number of retrieved relevant images to the total number of relevant images in the whole database.

The performance and efficiency measurement of all the techniques were computed to give a brief comparison between the systems through Precision and recall. Precision and recall is given as follows:

$$\text{Precision } (\text{P}) = \frac{\text{Total number of retrieved relevant images}}{\text{Total number of retrieved images}}$$

$$\text{Recall}(R) = \frac{\text{Total number of retrieved relevant images}}{\text{Total number ofrelevant images in the database}}$$

$$\text{Accuracy= (Precision + Recall) / 2}$$

As the whole database was known, every image of the database was used as query image. For each query image both precision and recall values were obtained. Finally the average value of precision and recall were obtained for each category of images of the database for all techniques and results for both Euclidean distance and Canberra distance are shown below in table 1 and table 2. The average percentage of accuracy of each method for two similar measurement metrices are shown in table 3 and for each category of images are shows in table 4 and table 5.

Table 1. Average Percentage of Recall for all methods

Distance Matrices	Color Based Search (HSV)	Color Based Search (YCbCr)	Color and wavelet Based Search (Symtel)	Color and wavelet Based Search (Daubechies)
Euclidean distance	61	64.4	70.9	71.66
Canberra distance	56.58	59	63.58	66.25

Table 2. Average Percentage of Precision for all methods

Distance Matrices	Color Based Search (HSV)	Color Based Search (YCbCr)	Color and wavelet Based Search (Symtel)	Color and wavelet Based Search (Daubechies)
Euclidean distance	72.66	76.66	84	84.83
Canberra distance	70	74.25	75.83	78.41

Table 3. Average Percentage of Accuracy all methods

Distance Matrices	Color Based Search (HSV)	Color Based Search (YCbCr)	Color and wavelet Based Search (Symtel)	Color and wavelet Based Search (Daubechies)
Euclidean distance	66.3	70.5	76.62	78.25
Canberra distance	64.5	68.37	69.93	71.2

Table 4. Percentage of accuracy for each category of images for Euclidean distance

Image Category	Color Based Search (HSV)	Color Based Search (YCbCr)	Color and Wavelet Based Search (Symtel)	Color and Wavelet Based Search (Daubechies)
Red Rose	84	70.5	71.5	72.5
Bird	55	66.5	70.5	74.5
Sunflower	41	83	89.5	93.5
Girl	76	72.5	69.5	72.5
Aquarium	80	68.5	95	95
Car	59	72.5	95	95
Girl in rain	69.5	34.5	64	64
Starfish	85.5	64	87.5	87.5
Baby	49.5	84	86.5	86.5
Tiger	77	78.5	78.5	78.5
Horse	48.5	73.5	54.5	57
Oranges	77	78.5	57.5	62.5

Table 5. Percentage of accuracy for each category of images for Canberra distance

Image Category	Color Based Search (HSV)	Color Based Search (YCbCr)	Color and Wavelet Based Search (Symtel)	Color and Wavelet Based Search (Daubechies)
Red Rose	81	70.5	70.5	71.5
Bird	55	64	60.3	61.5
Sunflower	61	64	66.5	68
Girl	73.5	74.5	79	79
Aquarium	74.5	70.5	77	80
Car	61	95	95	95
Girl in rain	61	38	41.2	42.6
Starfish	72.5	52.5	76	76
Baby	45	81	69.3	71
Tiger	45	74.5	77.5	79.5
Horse	76	64	64	66.5
Oranges	68.5	72	63.1	64.5

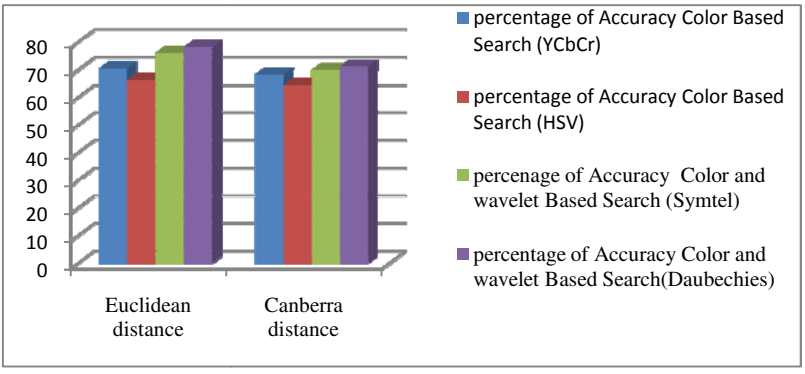


Fig. 3. Represents the table 3 in bar graphs

5 Conclusions

In this paper color histogram and wavelet transform were implemented and their performance was compared. The images having closest value compared to query image were retrieved from database as result. Then by taking some parameters the performance comparison between the two methods were obtained. The final results were shown through tables and graph. The complexity (cpu time) of wavelet based retrieval method were not significantly different than color based method because it take constant time for calculating the wavelet coefficients.

From all the experimental results, HSV color space based search gives an average retrieval accuracy of 66.3%, YCbCr color space based search gives an average retrieval accuracy of 70.5%, Symtel's wavelet based search gives an average retrieval accuracy of 76.62% and Daubechies's wavelet based search gives an average retrieval accuracy of 78.25% for Euclidean distance measurement (Table 3). Performance of Wavelet and color based search is better than the color based search because the former method extract the color feature as well as texture feature of image which gives additional information about the images.

References

[1] Brodatz, P.: Textures: A Photographic Album for Artists and Designers, 128 p. Dover Publications (1966)
[2] Stockman, G., Shapiro, L.: Computer Vision. Prentice Hall (2001)
[3] Li, H.C., Wei, L., Guo, H.L.: Research and Implementation of an Image Retrieval Algorithm Based on Multiple Dominant Colors. J. Comput. Res. Dev. 36, 96–100 (1999)
[4] Ali, A., Murtaza, S., Malik, A.S.: Content Based Image Retrieval Using Daubechies Wavelet Transform. In: Proceedings of the 2nd National Workshop on Trends in Information Technology, pp. 110–115 (2003)
[5] Vadivel, A., Majumdar, A.K., Sural, S.: Characteristics of Weighted Feature Vector in Content-Based Image Retrieval Applications. In: Proceedings of International Conference on Intelligent Sensing and Information Processing (IEEE Cat. No.04EX783), pp. 127–132 (2004)

[6] Vadivel, A., Majumdar, A.K., Sural, S.: Image Retrieval using Wavelet Based Texture Features. In: International Conference on Communications, Devices and Intelligent Systems, pp. 608–611 (2004)
[7] Suhasini, P.S., Krishna, K.S.R., Krishna, V.M.: CBIR Using Color Histogram Processing. J. Theor. Appl. Inf. Technol. 6, 116–122 (2009)
[8] Dubey, R., Choubey, R., Dubey, S.: Efficient Image Mining using Multi Feature Content Based Image Retrieval System. Int. J. Adv. Comput. Engg. Archit. 1, 17–25 (2011)
[9] Khan, W., Kumar, S., Gupta, N., Khan, N.: A Proposed Method for Image Retrieval using Histogram values and Texture Descriptor Analysis. Int. J. Soft. Comput. Eng. 1, 33–36 (2011)
[10] Khan, W., Kumar, S., Gupta, N., Khan, N.: Signature Based Approach For Image Retrieval Using Color Histogram and Wavelet Transform. Int. J. Soft. Comput. Engg. 1, 43–46 (2011)
[11] Sharma, N., Rawat, P., Singh, J.: Efficient CBIR Using Color Histogram Processing. Signal and Image Processing: An. Int. J. 2, 94–112 (2011)

A New Approach for Hand Gesture Based Interface

T.M. Bhruguram, Shany Jophin, M.S. Sheethal, and Priya Philip

Dept of Computer Science
Adi Shankara Institute of Engineering
and Technology, Kalady
shanyjophin.s@gmail.com

Abstract. This paper presents a new approach for controlling mouse movement and implementing mouse functions using a real-time camera. Most existing approaches involve changing mouse parts such as adding more but-tons or changing the position of the tracking ball. In-stead, we propose to change the hardware design. Our method is to use a camera, image comparison technology and motion detection technology to control mouse movement and implement its functions (right click, left click, scrolling and double click).

Keywords: HCI, Sixth Sense, VLCJ.

1 Introduction

As computer technology continues to develop, people have smaller and smaller electronic devices and want to use them ubiquitously. There is a need for new interfaces designed specifically for use with devices. Increasingly we are recognizing the importance of human computing interaction (HCI), and in particular vision-based gesture and object recognition. Simple interfaces already exist, such as embedded keyboard, folder-keyboard and mini-keyboard. However, these interfaces need some amount of space to use and can-not be used while moving. Touch screens are also a good control interface and nowadays it is used globally in many applications. However, touch screens can-not be applied to desktop systems because of cost and other hardware limitations. By applying vision technology and controlling the mouse by natural hand gestures, we can reduce the work space required. In this paper, we propose a novel approach that uses a video device to control the mouse system.

2 Related Work

2.1 Mouse Free

Vision-Based Human-Computer Interaction through Real-Time Hand Tracking and Gesture RecognitionVision-based interaction is an appealing option for

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 649–655.
springerlink.com

replacing primitive human-computer interaction (HCI) using a mouse or touch-pad. We propose a system for using a webcam to track a users hand and recognize gestures to initiate specific interactions. The contributions of our work will be to implement a system for hand tracking and simple gesture recognition in real time [1].

Many researchers in the human computer interaction and robotics fields have tried to control mouse movement using video devices. However, all of them used different methods to make a clicking event. One approach, by Erdem et al, used finger tip tracking to control the motion of the mouse. A click of the mouse buttonwas implemented by defining a screen such that a click occurred when a user's hand passed over the region [2, 3]. Another approach was developed by Chu-Feng Lien [4]. He used only the finger-tips to control the mouse cursor and click. His clicking method was based on image density, and required the user to hold the mouse cursor on the desired spot for a short period of time. Paul et al, used still another method to click. They used the motion of the thumb (from a 'thumbs-up' position to a fist) to mark a clicking event thumb. Movement of the hand while making a special hand sign moved the mouse pointer.

2.2 A Method for Controlling Mouse Movement Using a Real-Time Camera

This is a new approach for controlling mouse movement using a real-time camera. Most existing approaches involve changing mouse parts such as adding more buttons or changing the position of the tracking ball. Instead, we propose to change the hardware de-sign. Our method is to use a camera and computer vision technology, such as image segmentation and gesture recognition .Our method is to use a camera and computer vision technology, such as image segmentation and gesture recognition, to control mouse tasks (left and right clicking, double-clicking, and scrolling) and we show how it can perform everything current mouse devices can. This paper shows how to build this mouse control system [5].

2.3 Sixth Sense

'SixthSense' is a wearable gestural interface that augments the physical world around us with digital in-formation and lets us use natural hand gestures to interact with that information. The SixthSense prototype is comprised of a pocket projector, a mirror and a camera. The hardware components are coupled in a pendant like mobile wearable device. Both the projector and the camera are connected to the mobile computing device in the users pocket. The projector projects visual information enabling surfaces, walls and physical objects around us to be used as interfaces; while the camera recognizes and tracks user's hand gestures and physical objects using computer-vision based techniques [6].

2.4 Vlcj

The vlcj project is an Open Source project that pro-vides Java bindings for the excellent vlc media player from Video LAN. The bindings can be used to build media player client and server software using Java - everything from simply playing local media les to a full-blown video-on-demand streaming server is possible .vlcj is being used in diverse applications, helping to provide video capabilities to software in use on oceanographic research vessels and bespoke IPTV and home cinema solutions .vlcj is also being used to create software for an Open Source video camera at Elphel and video mapping for the Open Street Map project [7].

2.5 Mouseless

Mouseless is an invisible computer mouse that provides the familiarity of interaction of a physical mouse without actually needing a real hardware mouse. The Mouseless invention removes the requirement of having a physical mouse altogether but still provides the intuitive interaction of a physical mouse that we are familiar with. Mouseless consists of an Infrared (IR) laser beam (with line cap) and an Infrared camera. Both IR laser and IR camera are embedded in the computer. The laser beam module is modified with a line cap and placed such that it creates a plane of IR laser just above the surface the computer sits on. The user cups their hand, as if a physical mouse was present underneath, and the laser beam lights up the hand which is in contact with the surface. The IR camera detects those bright IR blobs using computer vision. The change in the position and arrangements of these blobs are interpreted as mouse cursor movement and mouse clicks. As the user moves their hand the cursor on screen moves accordingly. When the user taps their index finger, the size of the blob changes and the camera recognizes the intended mouse click.[8]

2.6 Real-Time Finger Tracking for Interaction

In this work, they described an approach for human finger motion and gesture detection using two cameras. The target of pointing on a flat monitor or screen is identified using image processing and line intersection. This is accomplished by processing above and side images of the hand. The system is able to track the finger movement without building the 3D model of the hand. Coordinates and movement of the finger in a live video feed can be taken to become the coordinates and movement of the mouse pointer for human-computer interaction purpose.[9]

3 Proposed Work

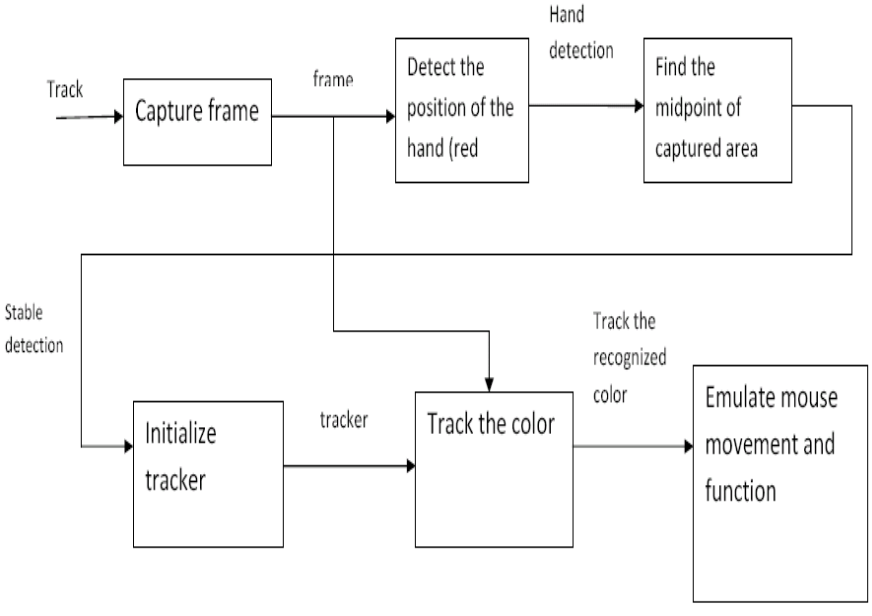


Fig. 1. The proposed working

3.1 Product and System Features

3.1.1 Computer

(Laptop)-Processor: Pentium4, Processor Speed: 1GHz, RAM Capacity: 512 MB, Hard disk: 40GB, Monitor: 15SVGA.

3.1.2 Webcam

It is used for the image processing and also for capturing the image.

Video data format: 12.24-bit RGB, Image resolution: Max 2560*2084,Software enhanced menu display/sec: 30 in CIF mode, Menu signal bit: 42db,Lens: 6.00mm,Vision: +/-28,Focus range: 3 centimeter to limitless.

3.1.3 Finger TIP

(Red and blue colored substance)- it is used as an alternative for mouse and control the functions of pointer.

3.1.4 Software Requirements

Operating System: Windows XP, Windows vista, Windows 7.

Code Behind: JAVA (Red Hat or Eclipse).

3.1.5 Internal Interface Requirements

Swing, vlcj.

3.2 Tools Used

Laptop (include software like red hat, eclipse, vlcj),webcam, a colored device.The performance of the software is made to be improved by examining each pixel leaving behind its four consecutive pixels.

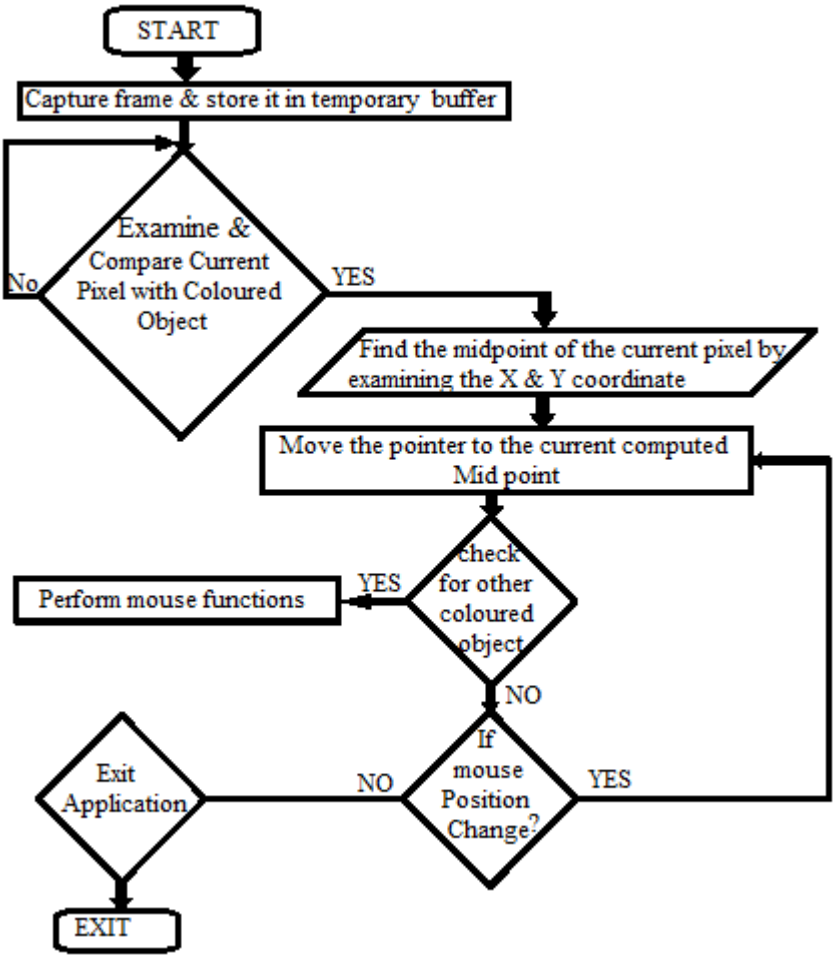


Fig. 2. The flow chart

4 Results

From our implementation and execution of our pro-gram we found that the mouse pointer can be made to move and its functions can be implemented without the use of a touchpad or mouse .The pointer is moved with the help of our finger gestures by

placing the specific color substance in our hand(any colored cap or any colored small substance) making us easy to use our system works. The performance of the software has been improved.

5 Future Work

There are still many improvements that can be made to our system like improving the performance of the current system and adding features such as enlarging and shrinking windows, closing window, etc. by using the palm and multiple fingers. The current system is variant to reflection and scale changes and requires proper hand gestures, good illumination technology and powerful camera for the performance of mouse functions. Precision can always be increased at the cost of recall by adding more stages, but each successive stage takes twice as much time to find harder negative samples and the applications which benefit from this technology. We present an image viewing application as an example of where this technology could lead to a more natural user interface. The same could be said for navigating something like Google Maps or browsing folders on a screen. But the applications reach far beyond that. They are particularly compelling in situations where touch screens are not applicable or less than ideal. For example, with projection systems there is no screen to touch. Here vision-based technology would provide an ideal replacement for touch screen technology. Similarly in public terminals, constant use results in the spread of dirt and germs. Vision-based systems would remove the need to touch such setups, and would result in improved interaction.

6 Conclusion

We developed a system to control the mouse cursor and implement its function using a real-time camera. We implemented mouse movement ,selection of the icons and its fuctions like right,left,double click and scrolling. This system is based on image comparison and motion detection technology to do mouse pointer movements and selection of icon. However, it is difficult to get stable results because of the variety of lighting and detection of the same colour in any other location in the background. . Most algorithms used have illumination issues. From the results, we can expect that if the algorithms can work in all environments then our system will work more efficiently. This system could be useful in presentations and to reduce work space. In the future, we plan to add more features such as enlarging and shrinking windows, closing window, etc. by using the palm and multiple fingers. The performance of the software can be only improved by small percentage due to the lack of a powerful camera and a separate processor for this application.

References

1. Park, H.: A Method for Controlling Mouse Movement using a Real-Time Camera, http://www.cs.brown.edu/research/pubs/theses/masters/2010/park.pdf2010 (online; acessed November 28, 2010)
2. Erdem, A., Yardimci, E., Atalay, Y., Cetin, V., Acoustics, A.E.: Computer vision based mouse. In: Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing, ICASS (2002)
3. Vision based Men-Machine Interaction, http://www.ceng.metu.edu.tr/~vbi/
4. Lien, C.-F.: Portable Vision-Based HCI - A Real-time Hand Mouse System on Handheld Devices
5. Jordan, B.C.: Hyokwon Lee mouse free (2009-2010), http://www.seas.upenn.edu/cse400/CSE400-2009-2010/nal-report/Jordan-lee.pdf (accessed November 28, 2010)
6. Pranavmistry, Sixthsense, http://www.youtube.com/watch?v=ZfV4R4x2SK0 (online; accessed November 28, 2010)
7. Mark dot lee Java bindings for the vlc media player, http://www.capricasoftware.co.uk/vlcj/index.php (online; accessed May 1, 2011)
8. Pranavmistry, Mouseless, http://www.pranavmistry.com/projects/mouseless/ (online; accessed November 28, 2010)
9. Shaker, N., Abou Zliekha, M.: Real-time Finger Tracking for Interaction. Damascus Univ., Damascus (2007), http://ieeexplore.ieee.org/search/freesrchabstract.jsp

Multi-document Summarization Based on Sentence Features and Frequent Itemsets

J. Jayabharathy[1], S. Kanmani[2], and Buvana[3]

[1] Assistant Professor, Department of Computer Science & Engineering, Pondicherry Engineering College, Puducherry, India
bharathyraja@pec.edu
[2] Professor, Department of Information Technology, Pondicherry Engineering College, Puducherry, India
[3] M.Tech Student, Department of Computer Science & Engineering, Pondicherry Engineering College, Puducherry, India

Abstract. Information retrieval is the process of searching for information and related knowledge within the collected documents or from the web Users and are presented with vast information which suffers from redundancy and irrelevance. Searching for the required information from this huge collection is a tiresome task. This motivated the researchers to provide high quality summary that allows the user to quickly locate the desired information. In this paper an attempt is made to improve the performance of summarization technique using the sentence features as length, position, centriod, Noun and by adding the new feature Noun-Verb pair. The second technique exploits modified FIS – Frequent Itemset Sequence generation algorithm for summarization. The redundancy elimination techniques are applied to achieve the efficient summary from various documents. The performance of proposed algorithms is compared with the existing MEAD summarization technique by considering F-measure. Introduction of Noun –Verb pair improves the quality of summarization compared to existing MEAD and our proposed FIS technique.

Keywords: Multi-document summarization, Query based summary, generic summary, frequent item set.

1 Introduction

Document Summarization is an automated technique, which reduces the size of the documents and gives the outline and concise information about the given document. Document summarization can be defined as a transformation process that consists in reducing the amount of information of a multimedia document and ends up with a simplified presentation of the initial content [25]. The process of summarization extracts the most important content from the batch of documents. In general, the summaries are created in two ways - Generic summary and Query based summary [20]. The generic summary refines overall content of the input document given by the user whereas the query based one retrieves the information that are more relevant to the user query. Document summarizations are of two types, they are single and multi-document

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 657–671.
springerlink.com

summarization. The summary that is extracted and created from a single document is known as Single Document Summarization, whereas Multi-document Summarization is an automatic procedure for the extraction of information from multiple sources. The purpose of a brief summary is to shorten the information search and to minimize the time by spotting the most relevant source documents. Summarizing and producing the concise information limits the need for accessing the original documents in some cases when fine tuning is required. Automated summaries give the extracted information from multiple sources algorithmically. Considering feature selection method to improves the summarization results [23]. Existing summarization technique [12] uses Feature profile considering word weight, sentence position, sentence length, sentence centrality, proper nouns in the sentence and numerical data in the sentence. Based on the feature profile sentence score is calculated for each sentence. The sentences with higher score are added to the summary. Our proposed work CPLNVN considers the features of [12] and in addition we have considered another feature called Noun – Verb pair which gives more meaningful information about the sentences. The sentence score is calculated for each and every sentence and the sentence score greater than the threshold is added to the summary. A new attempt is made to incorporate the Frequent Itemset Sequence generation method for summarization. FIS method is so far considered only for document clustering. We made an effort to include FIS based summarization of documents. This method uses the FIS generation algorithm for identifying the most frequent word set sequence in each sentences and support score is calculated. The sentences with support score greater than the threshold are added to the summary. According to different compression rates sentences are extracted from each cluster and ranked in the order of importance based on sentence score. F-measure is the performance metric which is used to compare the performance of our proposed algorithm with existing MEAD technique.

The remainder of this paper is organized as follows: Section 2 outlines the classification of various existing summarization techniques and describes about the related works in the field of generic based and query based summary generation. The general framework for extracting summary from document sources is discussed in section 3.The proposed summarization algorithms are described in section 4. Section 5 gives the detailed discussion about the experiments and gives detailed analyzes about the experimental results. The paper is concluded with future work in section 6.

2 Existing Work

This section gives an overview about various summarization techniques. The summarization techniques are classified into two major groups Generic and Query based summary creation. The generic summary refines overall content of the input document given by the user whereas the query based one retrieves the information that is more relevant to the user query. Some of the generic and query based summarization techniques are discussed below.

2.1 Generic Summary Extraction Technique

The RANDOM technique [9] is the simplest technique, which randomly selects lines from the input source documents. Depending upon the compression rate i.e. the size of

the summary, the randomly selected lines will be included to the summary. In this technique, a random value between 0 and 1 is assigned to each sentence of the document. A threshold value for length of the sentence is provided in general. The score of 0 to 1 is assigned to all sentences that do not meet assigned length cut-off. Finally, required sentences are chosen according to assigned highest score for desired summary.

LEAD[9] based technique is a one where first or first and last sentence of the paragraph are chosen depending upon the compression rate (CR) and it is suitable for news articles. The n% sentences are chosen from beginning of the text e.g. selecting the first sentence in the entire document, then the second sentence of each, etc. until the desired summary is constructed. In this technique a score of 1/n to each sentence is assigned, where n is the sentence number in the corresponding document file. This means that the first sentence in each document will have the same score; the second sentence in each document will have the same score, and so on. The length value is also provided as a threshold .The sentences with less length than the specified threshold value are ignored.

Dragomir R. Radev [1] et al proposed a multi-document text summarizer, called MEAD. The proposed system creates the summary based on cluster centroids. Centriod is the set of words that are most important to the cluster. In addition to the Centriod, position and first sentence overlap values are involved in the score calculation. Two new techniques namely cluster based relative utility and cross sentence information subsumption were applied to the evaluation of both single and multiple document summaries. Cluster base relative utility refers to the degree of relevance of a particular sentence to the general topic of the cluster. Summarization evaluation methods used could be divided into two categories: intrinsic and extrinsic. Intrinsic evaluation method measures the quality of multidocument summaries in a direct manner. Extrinsic evaluation methods measure how successfully the summaries help in performing a particular task. The extrinsic evaluation in terms called task-based evaluation. The new utility-based technique called CBSU was used for the evaluation of MEAD and of summarizers in general. It was found that MEAD produces summaries that are similar in quality to the ones produced by humans. MEAD's performance was compared to an alternative method, multi-document lead and showed how MEAD's sentence scoring weights can be modified to produce summaries significantly better than the alternatives.

MEAD is a commonly used technique which can perform many different summarization tasks. It can also summarize individual documents or clusters of related documents. MEAD is the combination of two baseline summarizers: lead-based and random based. Lead-based summaries generation is discussed in the previous paragraph. A random summary consists of enough randomly selected sentences (from the cluster) to produce a summary of the desired size. MEAD is a centriod-based extractive summarizer that scores sentences based on sentence level and inter-sentence features that indicates the quality of the sentence as a summary sentence. It then chooses the top-ranked sentences for inclusion in the output summary. MEAD extractive summaries score the sentences according to certain sentence features – Centriod [9], Position [9], and Length [9].

Afnan Ullah Khan [3] et al proposed a new technique for information summarization, which is the combination of the rhetorical structure theory and MEAD summarizer. In general MEAD summarizer is totally based on mathematical

calculation and lack a knowledge base. Rhetorical structure theory is used to overcome this weakness. The new summarizer system is evaluated against the original MEAD summarizer system. The proposed summarizer tool was exploited mainly in two areas of information that are Financial Articles and PubMed abstracts.

Dingding wang and Tao Li [22] integrated document summarization techniques into an incremental hierarchical clustering framework to re- organize sentence clusters immediately after new documents/sentences arrive so that the corresponding summaries can be updated efficiently. The hierarchical relationship among the sentences are displayed and re-constructed in real time. Shuzhi Sam Ge et al [24] proposed a sentence ranking and clustering based summarization method that extracts essential sentences from a document. To discover central sentences, a weighted undirected graph that takes sentence similarities and the discourse relationship between sentences as the weights of edges is constructed for the given document. A graph-ranking algorithm is implemented to calculate the scores of sentences.

Rasim M. Alguliev et al [25] devised a new document summarization model via sentence extraction to simultaneously deal with these two concerns during sentence selection.The optimization problem is solved by incorporating Discrete Particle Swarm Optimization based on Estimation of Distribution Algorithm (DPSO-EDA). The experimental results shows that DPSO-EDA is a very promising algorithm.

2.2 Query Based Summary Techniques

Dragomir R. Radev [2] et al designed a prototype system called SNS, which is pronounced as “essence”. This mainly integrates natural language processing and information retrieval techniques in order to perform automatic customized summarization of search engine results. The proposed system actually retrieves documents related to an unrestricted user query and summarizes a subset of them as selected by the user Task-based extrinsic evaluation shown that the system is of reasonably high quality. Xiao [6] et al designed and proposed a system to automate the multi-document summarization. The proposed system retrieves the documents related to the query given by the user. The sentence score is calculated based on relevant value and informativeness value. These values are realized by word sentence overlap and semantic graph techniques. Then the sentences with the highest score are included to the summary. The investigational result achieves better quality.

3 General Architecture for Summarization

Usually document sources are of unstructured format, transforming these unstructured documents to structured format requires some pre-processing steps. Fig.1 presents the sequence of steps involved in document Summarization. Some commonly used pre-processing steps are given below.

3.1 Preprocessing Phase

Tokenization. The process of splitting the sentences into separate tokens. For example, “this is a paper about document summarization” is splitted as this\is\paper\ about\document\summarization.

Stop Words Removal. Stop words are typical frequently occurring words that have little or no discriminating power, such as \a", \about", \all", etc., or other domain-dependent words.

Stop words are often removed.

Stemming. Removes the affixes in the words and produces the root word known as the stem [13]. Typically, the stemming process is performed so that the words are transformed into their root form. For example connected, connecting and connection would be transformed into 'connect'. Most widely used stemming algorithms are Porter [17], Paice stemmer [16], Lovins [15], S-removal [14]

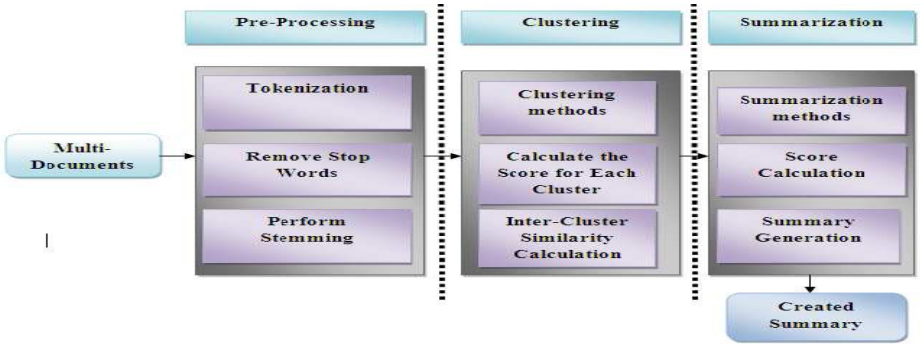


Fig. 1. General architecture for summarization

3.2 Clustering Phase

Feature Vector Construction. Feature vector is constructed based on term frequency (TF-DF) and inverse document frequency (TF-IDF). After applying the preprocessing techniques, the processed documents are clustered using a clustering algorithm in order to group the similar documents. Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations of same cluster are similar in some sense. Some of the popular types of clustering are described below.

Hierarchical algorithms find consecutive clusters using previous clusters. They are of two types namely agglomerative ("bottomup") and divisive ("top-down"). The first types begin with each element as an individual cluster and merge them into larger clusters. Divisive algorithms start with the whole document set and divide it into smaller clusters. Partitional algorithms typically resolve all clusters at once, but can be used as divisive algorithms in the hierarchical clustering [21].

We make use the FIHC (Frequent Itemset Based Hierarchical Clustering) [5] for clustering the given document sets. After the clustering process the summary is created for the clustered documents. The summarization process applied is discussed below.

3.3 Existing Summarization Technique

This technique [12] includes some of the characteristic like centriod, position, sentence length, noun and numerical data for the creation of the multi-document summary.

Score Calculation
For calculating the sentence score the following features are considered:

Centriod

Centriod is the set of words which are more important in the cluster. Centriod value is calculated for each and every sentence separately. For calculating the centriod value the following formula is used.

$$Centroid = C_i = \sum \frac{(TF - IDF(W_i))}{|D|} \tag{1}$$

Where, TF represents the term frequency of the particular word in a sentence and IDF represents the inverse document frequency of word.TF-IDF which is used to calculate the importance of the term in a particular document. D represents the total number of documents in a particular cluster for which the summary is to be created. The summation of these values gives the centriod value of each and every sentence in a document which belongs to particular cluster.

Position

The positional value gives the position of the particular sentence in a document. For example first sentence will get the positional value of 1. The positional value is calculated by using the formula mentioned below.

$$Position = Pi = \frac{n - i + 1}{n} * \max(centroid) \tag{2}$$

Where "n" represents the total no of sentences in a document and "i" sentence number between 1 to n.

Sentence Length

The length of the sentence plays an important role in calculating the importance of the sentence. The length of the sentence is calculated by using the formula which is described below.

$$Length = L_i = \frac{|S_i|}{\max(S_z) \in D(S_i)} \tag{3}$$

Where $|Si|$ represents the number of characters of sentence i[th] sentence and max $Sz \in D(Si)$ represents the maximum number of characters in a sentence that belongs to $D(Si)$.

Proper Noun Count

The sentences with more nouns are considered to be important and are added to summary. The formula for calculating the proper noun score is

$$Noun = N_i = \frac{\sum (N(S_i))}{L_i} \tag{4}$$

Where L_i is the length of the sentence and $N(S_i)$ is number of the nouns in the sentence.

Numerical Data Count

The sentence which contains the numerical data is considered to be important in calculating the sentence score. Numerical information in news articles, Cricket data sources and Medical Document collections are considered to more informative than any other data. The numerical data value of the particular sentence is calculated by the formula which is described below. Where, $\sum ND_i$ represents the total count of all numerical data in a particular sentence and L_i represents the length value of the particular sentence.

$$Numerical\ data = ND_i = \frac{\sum ND_i}{L_i} \tag{5}$$

Then for each and every cluster score is calculated by using the above said features. Finally the sentences inside each and every document are arranged in the ascending orders of the scored sentences are included in the summary and the process continues until the compression ratio is met.

4 Proposed Work

4.1 Proposed Summarization Technique Based on CPLNVN

Our proposed technique includes the characteristic like centriod, position, sentence length, noun, verb and numerical values for the creation of the multi-document summary. Centroid, Position and Sentence length, numerical data count are calculated using the existing formulas. In addition to that we have included Noun-Verb importance calculation since noun verb pair gives significant information about the sentences.

Noun and Verb Count

In practice the sentence which contains more proper verb and noun is more important and those sentences are included in the summary most probably. The formula which is used to calculate the noun and verb is described below.

$$Noun\ Verb = NVi = \frac{\sum N(S_i) + V(S_i)}{L_i} \tag{6}$$

Where $N(S_i)$ and $V(S_i)$ are the nouns and verbs in the sentence S_i and L_i is the length of the sentence. A sentence may contain more than one noun and verb. This gives the frequency of nouns and verbs in the sentences of the whole document. The Nouns and verbs are extracted from each sentence by using the tool named "Stanford Tagger".

Algorithm design of the proposed CPLNVN summarization method:

1. Given the clustered document set 'Ci', Where, i=1, 2…….n
2. For each document form compute Ci
3. Calculate the sentence score by using various characteristic like Centriod, Position, Length, noun verb and numerical data value.

$$Centroid = C_i = \sum \frac{(TF - IDF(W_i))}{|D|}$$

$$Length = L_i = \frac{|S_i|}{\max(S_z) \in D(S_i)}$$

$$Position = Pi = \frac{n-i+1}{n} * \max(centroid)$$

$$Noun\ Verb\ = NVi\ = \frac{\sum N(S_i) + V(S_i)}{L_i}$$

$$Numerical\ data = NDi = \frac{\sum ND_i}{L_i}$$

4. Add the characteristic value for Sentence S_i, where i= {1,2……n} to get the total score of the sentence and store it in an array in descending order.
5. Add sentence that has the highest score to the summary.
6. Select sentence that has the highest score in rest of the sentences;
7. Get the redundancy penalty (R_s) for each sentence which overlaps with sentences that have high score values - redundancy value is obtained by calculating the overlap value for all sentences with the second highest scored sentence which is selected in step 6.
   R_s=2*(# overlapping words) / (#words in sentence1 + #words in sentence2).
   R_s=1, when the sentences are identical and R_s=0 when they have no words in common.
8. Subtract the redundancy penalty score from the original score of the sentence and include the sentences in an array according to their score value.
9. Include the sentences in the summary one by one from the array until the compression ratio is met.

Sample feature value calculation for the following paragraph

Cricket is a bat-and-ball team game. Many variations exist, with its most popular form played on an oval shaped outdoor arena known as a cricket field at the centre of which is a rectangular 22-yard long pitch that is the focus of the game. A game is contested between two teams of eleven players each.

Centroid value

Cricket is a bat-and-ball team game.
Tf-idf(cricket)=0.34 Tf-idf(ball)=0.22
Tf-idf(bat)=0.22
Toal no documents=1 Centroid=0.78

Many variations exist, with its most popular form played on an oval shaped outdoor arena known as a cricket field at the centre of which is a rectangular 22-yard long pitch that is the focus of the game.

Tf-idf(variation)=0.22	Tf-idf(exist)=0.22	Tf-idf(centre)= 0.22
Tf-idf(popular)=0.22	Tf-idf(play)=0.22	Tf-idf(rectangular)= 0.22
Tf-idf(oval)=0.22	Tf-idf(shape)=0.22	Tf-idf(yard)= 0.22
Tf-idf(outdoor)= 0.22	Tf-idf(arena)= 0.22	Tf-idf(long)= 0.22
Tf-idf(cricket)= 0.34	Tf-idf(field)= 0.22	Tf-idf(pitch)= 0.22
Tf-ifd(focus)= 0.22	f-idf(game)=0.34	

Total no documents=1
Centroid=3.98

A game is contested between two teams of eleven players each.
For above sentence the Centroid is 1.44

Position value
Cricket is a bat-and-ball team game=((3-1+1)/3)*0.34=0.34
Many variations exist, with its most popular form played on an oval shaped outdoor arena known as a cricket field at the centre of which is a rectangular 22-yard long pitch that is the focus of the game=((3-2+1)/3)*0.34=0.226
A game is contested between two teams of eleven players each=((3-3+1)/3)* 0.34=0.113

Sentence length value

Cricket is a bat-and-ball team game=28/157=0.178
Many variations exist, with its most popular form played on an oval shaped outdoor arena known as a cricket field at the centre of which is a rectangular 22-yard long pitch that is the focus of the game=157/157=1
A game is contested between two teams of eleven players each=50/157=0.318

Noun verb count value

Cricket/NNP is/VBZ a/DTbat-and-ball/JJ team/NN game/NN =3/0.178=16.85

Many/JJ variations/NNS exist/VBP ,/, with/IN its/PRP$ most/RBS popular/JJ form/NN played/VBN

on/IN an/DT shaped/JJ outdoor/JJ arena/NN known/VBN as/INa/DT cricket/NN field/NN at/IN the/DT center/NN of/IN which/WDT is/VBZ a/DT rectangular/JJ 22-yard/JJ long/JJ pitch/NN that/WDT

is/VBZ the/DT focus/NN of/IN the/DT game/NN = 8/1=8

A/DT game/NN is/VBZ contested/VBN between/IN two/CD teams/NNS of/IN eleven/NN players/NNS

each/DT =2/0.318=6.28

Numerical data value

Cricket is a bat-and-ball team game=0

Many variations exist, with its most popular form played on an oval shaped outdoor arena known as a cricket field at the centre of which is a rectangular 22-yard long pitch that is the focus of the game=1

A game is contested between two teams of eleven players each=0

4.2 Proposed - FIS (Frequent Item Set) for Summarization

Frequent item set methods identify the frequent word sequence from the clustered document by using the support count value. Always the minimum support value is given by the human. A set of words that appear together in more than one minimum fraction of documents is called itemset and item is one which belongs to one frequent itemset. An item or itemset which exceeds the minimum support value will be selected along with the sentence and included in the summary. The sentences are included in the summary until the compression ratio is met.

CFWS[19] algorithm is used for generating the Frequent Itemset from the documents. In this paper the clustering process is based on the Frequent Itemset generation. We made an attempt to use this FIS generation algorithm for document summarization.

Algorithm design of the proposed FIS based summarization method

1. Given the clustered document set ‘Ci’ Where, i=1, 2…….n
2. From each document form Ci find all frequent itemset based on CFWS algorithm
   //Get frequent items: Items whose occurrence in database is greater than or equal to the min.support threshold.
   //Get frequent itemsets:
   Generate candidates from frequent items and Prune the results to find the frequent itemsets.
3. Scan the transaction database to get the support S of each item.
4. If S>=min Support, then add to Frequent 1-itemsets, L1.

5. Use L_{k-1} join L_{k-1} to generate a set of candidate K-itemsets.
6. Scan the transaction database to get the support S of each candidate K-itemset.
7. If S>=min Support then, add to K-Frequent Itemsets.
8. Include the sentences in the summary which contains the itemset which is found in the step7.
9. Repeat step 8 until the compression ratio is met or the desired percentage is achieved.

5 Experimental Results

A corpus of 100 cricket documents was selected from Cricket websites and also from the Wikipedia articles. Then these 100 documents are clustered using FIHC[5] in order to group the similar documents in a text file. Then the summary is created based on existing method MEAD and also for the same corpus we have created the summary using our proposed CPLNVN and FIS methods. Then the created summary is evaluated using FMeasure as the performance metric. F-Measure is computed using the formula given below.

$$\text{F-Measure}=\frac{(2*(\text{Pr}\,ecision*\text{Re}\,call))}{\text{Pr}\,ecision+\text{Re}\,call} \tag{7}$$

Where, Precision reflects correctness of number of systems extracted sentences Recall reflects number of missed correct missed sentences by the system. F-Measure ranges from 0 to 1.

The system is refined based on the values of these parameters. The precision is measured by using the following parameters.

- **Correct**-the number of sentences extracted by the system as well as by the human
- **Wrong**-the number of sentences extracted by the system but not by the human
- **Missed**-the number of sentences extracted by the human but not by the system

Using the above parameters, we calculate $\text{Precision}=\frac{correct}{(correct+wrong)}$ (8)

$$\text{Recall}=\frac{correct}{(correct+missed)} \tag{9}$$

Fig. 2. shows the Precision value for the MEAD, CPLNVN and FIS. The Precision value for the MEAD ranges from 0.632 to 0.770, CPLNVN ranges from 0.986 to 1 and FIS ranges from 0.84 to 1. Fig 3 shows the Recall value for the MEAD, CPLNVN and FIS. The Recall value for MEAD ranges from 0.653 to 0.714, CPLNVN ranges from 0.90 to 0.999 and for FIS it ranges from 0.83 to 1. Fig 4 shows the F-Measure value for the MEAD, CPLNVN and FIS. The F-Measure value for the MEAD ranges from 0.665 to 0.728, CPLNVN ranges from 0.947 to 0.997 and FIS ranges from 0.834 to 0.947.

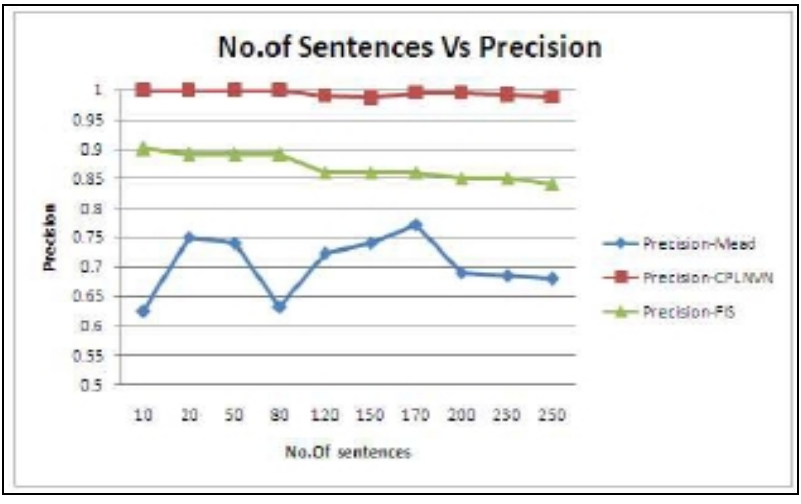


Fig. 2. No of sentences Vs Precision

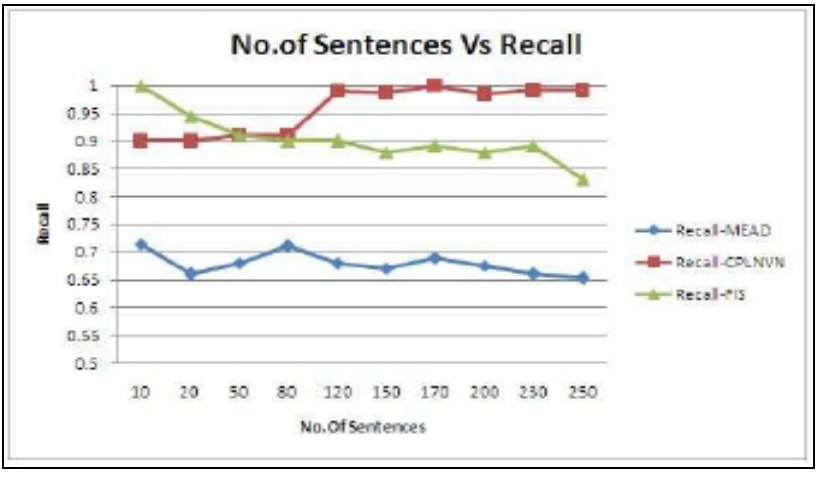


Fig. 3. No of sentences Vs Recall

From these results it is inferred that both of the proposed methods achieve higher F-Measure when compared existing MEAD method. The average percentage of improvement in F- measure quality for CPLNVN and FIS is +28% and +19% compared to MEAD technique. FIS method generates efficient summary when the number of sentences are very less because, Item set are identified accurately when the sentences are less in number. Irrespective of the size and count of documents our CPLNVN produces well-organized summary compared to FIS (proposed work II) and MEAD (existing method). It is proved that the proposed CPLNVN and FIS generate more precise and efficient summary when compared to the existing method.

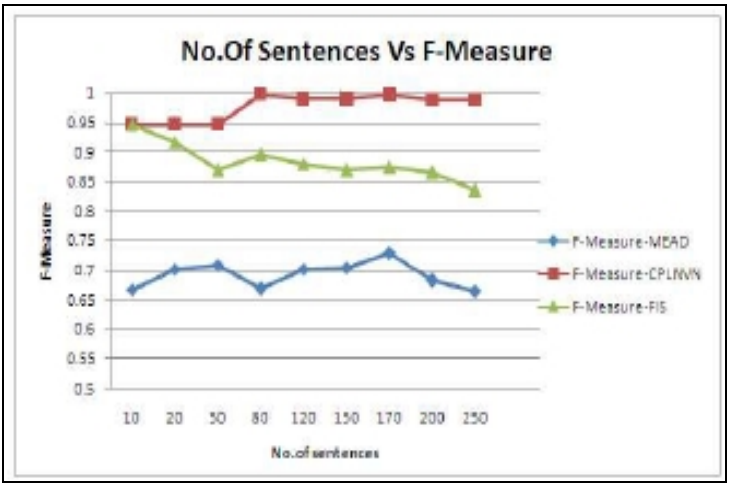


Fig. 4. No of sentences Vs F-Measure

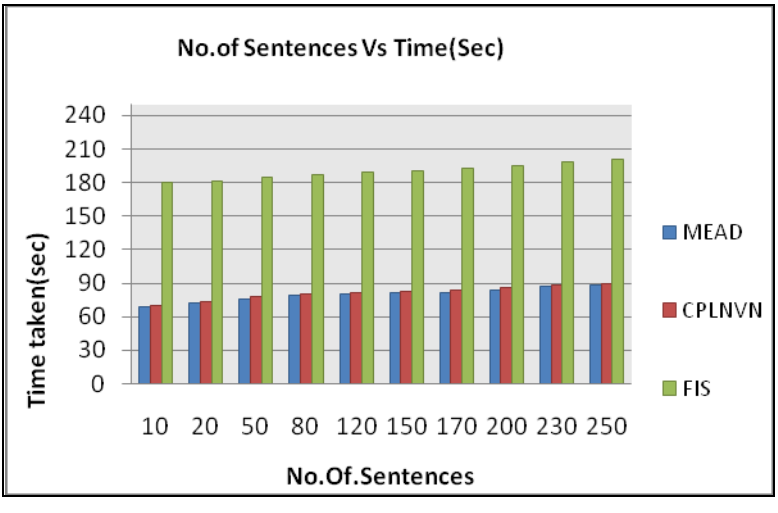


Fig. 5. No of sentences Vs Time Taken for computation

Fig 5 illustrates the comparison of time computation for the existing and proposed algorithms. Though our proposed FIS algorithms takes more time for computation compared to existing MEAD technique, the quality of summary generated by this method is more concise and relevant to the input documents, whereas the computation time taken by CPLNVN method is approximately equivalent to MEAD but quality wise our method achieves better improvement. As the summaries generated are for static document clusters, time taken for computation is not an important issue. Once the summaries are generated, it is appropriate until there are any updates in the formulated clusters.

6 Conclusions

The multi-document summarization has high demand in today's world because of the information overload. Information is available in various formats from various sources. To infer all the information in a shorter period is tiresome task and also the user wants the information to be more precise and quickly readable. We have proposed two summarization techniques namely CPLNVN and FIS. CPLNVN technique creates the summary based on Centriod, Position, Sentence Length, Noun Verb and Numerical Data. The second work considers the frequent itemset based summarization concept. The sentence with highest score is included in the summary until the compression ratio is met. The proposed methods are compared with existing MEAD technique using F-Measure as the performance metric. The results prove that the proposed algorithm produces concise and efficient summary compare to MEAD.

Future enhancements

The system can be enhanced to create dynamic summary in the distributed environment and it could be enhanced to summarize not only text documents but also other type of documents like PDF etc. Instead using TF-IDF for document representation, semantic concepts may be used for clustering and summarization.

References

1. Radev, D.R., Jing, H., Stys, M., Tam, D.: Centriod based summarization of multiple documents. Information Processing and Management 6, 869–1038 (2004)
2. Radev, D.R., Fan, W.: Automatic summarization of search engine hit lists. In: Proceedings of the ACL 2000 Workshop on Recent Advances in Natural Language Processing and Information Retrieval, vol. 11, pp. 99–109 (2000)
3. Khan, A.U., Khan, S., Mahmood, W.: MRST:A NewTechnique for Information Summarization. Proceedings of World Academy of Science, Engineering and Technology 4, 249–255 (2005)
4. Zhang, S., Zhao, T., Zheng, D., Zhao, H.: Two stage sentence selection approach for multi-Document summarization. Journal of Electronics 2(4), 562–567 (2008)
5. Wei, F., He, Y., Li, W., Lu, Q.: A Query-Sensitive Graph-Based Sentence Ranking Algorithm for Query-Oriented Multi-Document Summarization. In: International Symposiums on Information Processing, pp. 9–13 (2008)
6. Yang, X.-P., Liu, X.-R.: Personalized Multi-Document Summarization in Information Retrieval. In: Seventh International Conference on Machine Learning and Cybernetics, Kunming, July 12-15, pp. 4108–4112 (2008)
7. Wang, D., Li, T., Zhu, S., Ding, C.: Multi-Document Summarization via Sentence–Level Semantic Analysis and Symmetric Matrix Factorization. In: SIGIR, Singapore, July 20-24, pp. 307–314 (2008)
8. Hachey, B.: Multi-Document Summarization Using Generic Relation Extraction. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 420–429 (2009)

9. Ali, M.M., Ghosh, M.K., Al-Mamun, A.: Multi-document Text Summarization: Sim With First Based Features and Sentence Co-selection Based Evaluation. In: International Conference on Future Computer and Communication, April 3-5, vol. 12, pp. 93–96 (2009) ISBN–13: 978-0-7695-3591-3
10. Huang, L., He, Y., Wei, F., Li, W.: Modeling Document Summarization as Multi-objective Optimization. In: Third International Symposium on Intelligent Information Technology and Security Informatics, April 2-4, pp. 382–386 (2010)
11. Gong, S., Qu, Y., Tian, S.: Subtopic-based Multidocuments Summarization. In: Third International Joint Conference on Computational Science and Optimization, pp. 382–386 (2010)
12. Kogilavani, A., Balasubramani, P.: Clustering and Feature Specific Sentence Extraction Based Summarization of Multiple Documents. International Journal of Computer Science & Information Technology (IJCSIT) 2(4), 99–111 (2010)
13. Frakes, W.B., Fox, C.J.: Strength and Similarity of Affix Removal Stemming Algorithms. In: ACM SIGIR Forum, pp. 26–30 (2003)
14. Harman, D.: How Effective is Suffixing. Journal of the American Society for Information Science 42(1), 7–15 (1991)
15. Lovins, J.B.: Development of a Stemming Algorithm. Mechanical Translation and Computational Linguistic 11, 22–31 (1968)
16. Paice, C.D.: Another Stemmer. In: SIGIR Forum, vol. 24(3), pp. 56–61 (1990)
17. Porter, M.F.: An Algorithm for Suffix Stripping. Program. 14, 130–137 (1980)
18. Fung, B., Wnag, K., Ester, M.: Hierarchical Document Clustering using Frequent itemsets. In: SIAM International Conference on Data Mining, SDM 2003, pp. 59–70 (2003)
19. Li, Y., Chung, S.M., Holt, J.D.: Text document clustering based on frequent word meaning sequences. Journal, of Data & Knowledge Engineering 64(1), 381–404 (2008)
20. http://en.wikipedia.org/wiki/Document_summarization
21. Prathima, Y., Supreethi, K.P.: A Survey Paper on Concept Based Text Clustering. International Journal of Research in IT & Management 1(3), 45–60 (2011)
22. Wang, D., Li, T.: Document Update Summarization using Incremental Hierarchical Clustering. In: Proceedings of the Conference on Information and Knowledge Management (CIKM 2010), pp. 279–288 (October 2010)
23. Xiong, Y., Liu, H., Li, L.: Multi-Document Summarization Based on Improved Features and Clustering. In: IEEE International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE), pp. 1–5 (August 2010)
24. Ge, S.S., Zhang, Z., He, H.: Weighted Graph Model Based Sentence Clustering and Ranking for Document Summarization. In: 4th International Conference on Interaction Sciences (ICIS), pp. 90–95 (August 2011)
25. Alguliev, R.M., Aliguliyev, R.M., Mehdiyev, C.A.: An Optimization Model and DPSO–EDA for Document Summarization I. J. of Information Technology and Computer Science (5), 59–68 (2011)

Performance Evaluation of Evolutionary and Artificial Neural Network Based Classifiers in Diversity of Datasets

Pardeep Kumar[1], Nitin[1], Vivek Kumar Sehgal[1], and Durg Singh Chauhan[2]

[1] Department of CSE & ICT, Jaypee University of Information Technology, Waknaghat, Solan-173234, Himachal Pradesh, India
[2] Uttrakhand Technical University, Dehradun, Uttarakhand, India
pardeepkumarkhokhar@gmail.com,
{delnitin,vivekseh}@ieee.org,pdschauhan@acm.org

Abstract. In the last two decades, we have seen an explosive growth in our capabilities to both generate and collect data. Advances in scientific data collections (e.g. from remote sensors or from space satellites), the widespread use of bar codes for almost all commercial products, and the computerization of many business and government transactions have generated a sea of data. So there is a need for automatic tools and techniques for such a huge collection of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD) and data mining. Data mining plays an important role to discover important information to help in decision making of a decision support system. It has been the active area of research in the last decade. The classification is one of the important tasks of data mining. Different kind of classifiers have been suggested and tested to predict the future events based on unseen data. This paper compares the performance evaluation of evolutionary based genetic algorithm and artificial neural network based classifiers in diversity of datasets. The performance evaluation metrics are predictive accuracy, training time and comprehensibility. Evolutionary based classifier shows better comprehensibility and predictive accuracy as compared to ANN based classifier. Such a classifier is slower as compared to the ANN based one.

Keywords: Knowledge Discovery in Databases, Evolutionary Computation, Artificial Neural Network, STAGLOG.

1 Introduction

Information plays a vital role in business organizations. Today's business is information hungry. Information can be used by the top level management for decision making to make future policies. Due to increasing size of organizations data rapidly, manual interpretation of data for information discovery is not feasible.

This never ending cycle of data generation followed by the increased need to store that data has been a challenge faced by information systems professionals for decades. Despite successes in recent years in the area of large scale database design, we are still challenged by the difficulties associated with unlocking the data we need and

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 673–682.
springerlink.com

removing it from the cavernous databases in which it resides. In addition, we are becoming increasingly aware of the hidden treasure trove of new knowledge quietly residing in our data and face considerable frustrations when we attempt to get it. This constant cycle of data generation, storage and difficulty in retrieval and analysis has resulted in the development of new and powerful tools to assist us in meeting this challenge.

Over the last three decades, data mining has been growing on the map of computer science. It deals with the discovery of hidden knowledge, unexpected patterns and new rules from large databases. Data mining is regarded as the key element of a much more elaborate process called Knowledge Discovery in Databases (KDD) which is defined as the non–trivial process of identifying valid, novel, and ultimately understandable patterns in large databases [1]. One of the important tasks of data mining is classification. The conventional classifiers used for classification are decision trees, neural network, statistical and clustering techniques. There is lot of research going in the machine learning and statistics communities on classifiers for classification. In the recent past, there has been an increasing interest in applying evolutionary methods to Knowledge Discovery in Databases (KDD) and a number of successful applications of Genetic Algorithms (GA) and Genetic Programming (GP) to KDD have been demonstrated.

A study, called the STATLOG Project compares the predictive accuracy of several decision tree classifiers on a large number of datasets [2]. The STATLOG Project finds that no classifier is uniformly most accurate over the datasets studied and many classifiers possess comparable accuracy. Earlier comparative studies put emphasis on the predictive accuracy of classifiers; other factors like comprehensibility and classification index are also becoming important. Breslow and Aha have surveyed methods of decision tree simplification to improve their comprehensibility [3]. Brodley and Utgoff , Brown, Corruble, and Pittard, Curram and Mingers, and Shavlik Mooney and Towell have also done comparative studies in the domain of classifiers [4-7]. Saroj and K.K Bhardwaj have done excellent work on GA's ability to discover production rules and censor based production rules [8]. Earlier comparative studies put emphasis on the predictive accuracy of classifiers; other factors like comprehensibility and classification index are also becoming important.

This paper compares evolutionary approach based genetic algorithm and ANN (Back Propagation) based classifiers. These classifiers are tested on four datasets (Mushroom, Vote, Nursery and Credit) that are taken from the University of California Irvine, Repository of Machine Learning Databases (UCI) [9]. Here, section 2 briefly describes the classifiers and section 3 describes some background to the datasets and experimental setup, and Section 4 shows the result. Conclusion is given in section 5.

2 The Classifiers

2.1 Artificial Neural Network

Back Propagation learns by iteratively processing a set of training samples, comparing the network's prediction for each sample with the actual known class label. For each

training sample, the weights are modified so as to minimize the mean squared error between the network's prediction and the actual class. These modifications are made in backward direction, that is, from the output layer, through each hidden layer down to the first hidden layer (hence the name back propagation). In general, the weights will eventually converge, and the learning process stops [10-12].

2.2 Genetic Algorithm

Genetic Algorithms (GAs) are based on Darwinian natura selection and Mendelian genetics, in which each point in the search space is a string called a chromosome that represents a possible solution. This approach requires a population of chromosomes representing a combination of features from the set of features and requires a cost function that calculates each chromosome's Fitness (this function is called evaluation function or Fitness function). The algorithm performs optimization by manipulating a finite population of chromosomes. In each generation, the GA creates a set of new chromosomes by crossover, inversion and mutation, which correlate to processes in natural reproduction [13-14]. Evolutionary approach based classifiers use fitness function in data mining is defined by

$$\begin{aligned}&\text{Fitness (Chromosome i. e. A rule defined in population)}\\&= \text{K x Predictive Accuracy x Comprehensibility}\end{aligned} \tag{1}$$

Pseudo code for genetic algorithm is given below

1. *Pseudo code1:*
2. *Create initial population;*
3. *Compute fitness of individuals;*
4. *REPEAT*
5. *Select individuals based on fitness;*
6. *Apply genetic operators to selected individuals, creating offspring;*
7. *Compute fitness of offspring;*
8. *Update the current population;*
9. *UNTILL (stopping criteria) [15-16].*
10. *In this pseudocode, initial population represents encoded production rules. Fitness function is defined in terms of predictive accuracy and comprehensibility. Fitness function is used in selection process while mining rules from the data set. Pseudo code for selection process when genetic algorithm is used in mining is given below*
11. *Pseudo code2:*
12. *Compute Fitness of each rule(individual);*
13. *(assume that the higher the fitness, the better the rule)*
14. *Sort rules in decreasing order of Fitness;*
15. *Store the sorted rules into CandidateRuleList;*
16. *WHILE (CandidateRuleList is not empty) AND (TrainingSet is not empty)*
17. *Remove from the TrainingSet the data instances correctly covered by the first rule in CandidateRuleList ;*
18. *Remove the first rule from CandidateRuleList and insert it into SelectedRuleList;*

3 Experimental Setup

3.1 Datasets

There are four datasets used in this research work from real domain. These datasets are available from UCI machine learning repository [8].

Mushroom: This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota family. Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one.

Vote: This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA. The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to yea), voted against, paired against, and announced against (these three simplified to nay), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition). The class attribute contains two values: Democrat and Republic**.**

Nursery: This data set was originally developed to rank applications for nursery schools. It was used during several years in 1980's when there was excessive enrollment to these schools in Ljubljana, Slovenia, and the rejected application frequently needed an objective explanation. This data set is used to predict whether application is rejected or accepted. The final decision depends on occupation of parents and child's nursery, family structure and financial standing, and social and health picture of the family. The class attribute contains five values: not_recom, recommend, very_recom, priority and spec_prior.

Credit: This data set concerns credit card application. Talking to the individuals at a Japanese company that grants credit generated the dataset. The class attribute represents positive and negative instances of people who were and were not granted credit. The class attribute is represented by +(Credit granted) and –(credit not granted). All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data.

3.2 Parameter for Comparison

1. Predictive accuracy: It is defined as the percentage of correct prediction made by a classifier [1, 10]. The formula for Predictive accuracy is calculated by the equation 2

$$P.A. = 100 * \frac{True\ pridiction}{Total\ Prediction} \tag{1}$$

2. Training time: It is defined as the time that a classifier takes to build a model on datasets. Minimum training time is desirable. [10-11]
3. Comprehensibility: It shows degree of simplicity in rule sets obtained after classification. Higher degree of comprehensibility is required. Greater the number of leaf nodes and depth of tree, lesser will be the comprehensibility [5, 11].

3.3 Implementation

ANN based back propagation classifier has been tested using Clementine 10.1 on a Pentium IV machine with Window XP platform. Evolutionary based GA classifier has been tested using GALIB 245 simulator on Linux platform.

4 Results

4.1 Predictive Accuracy

Table 1 and fig. 1 shows the predictive accuracy on Mushroom, Vote, Nursery and Credit datasets.

Table 1. Predictive Accuracy on Datasets

Data Set / Classifier	Mushroom	Vote	Nursery	Credit
Neural N/W	100%	92.21%	92.34%	86.88%
Genetic Algorithm	98%	94%	97.3%	96.2%

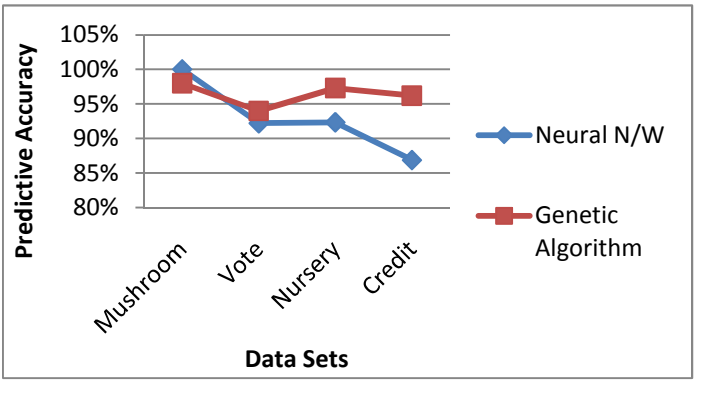


Fig. 1. Predictive accuracy on datasets

In fig. 1, evolutionary approach based genetic algorithm classifier shows good predictive accuracy as compared to ANN based classifier irrespective of the domains of the datasets.

4.2 Training Time

Table 2 and fig. 2 show the training time on Mushroom, Vote, Nursery and Credit data set.

Table 2. Training Time (in Sec.)on datasets

Data Set / Classifier	Mushroom	Vote	Nursery	Credit
Neural N/W	4 .	1	1	1
Genetic Algorithm	3	1	2	3

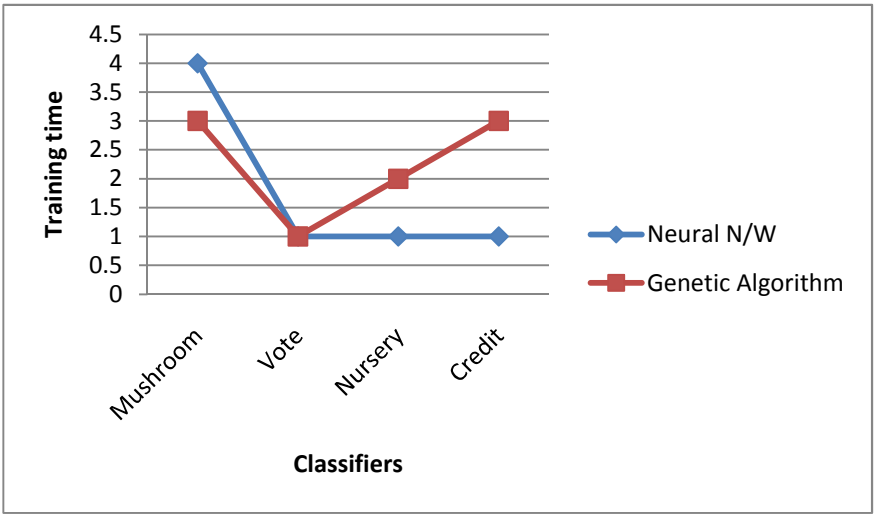


Fig. 2. Training time (Sec) on datasets

In fig. 2, evolutionary approach based genetic algorithm classifier shows higher training time as compared to ANN based back propagation classifier irrespective of the domains of the datasets.

4.3 Comprehensibility

Table 3 and Fig. 3-6 show the comprehensibility on Mushroom, Vote, Nursery and Credit datasets.

Table 3. Comprehensibility on datasets

Data Set / Classifier	Mushroom		Vote		Nursery		Credit	
	Leaf Node	Depth	Leaf Node	Depth	Leaf Node	Depth	Leaf Node	Depth
Neural N/W	Nil	Nil	Nil	Nil	Nil	Nil	Nil	Nil
Genetic Algorithm	6	4	8	6	4	2	5	3

In fig. 3-6, evolutionary approach based genetic algorithm classifier shows better comprehensibility as compared to ANN based back propagation classifier. The second one shows poor comprehensibility over all the datasets.

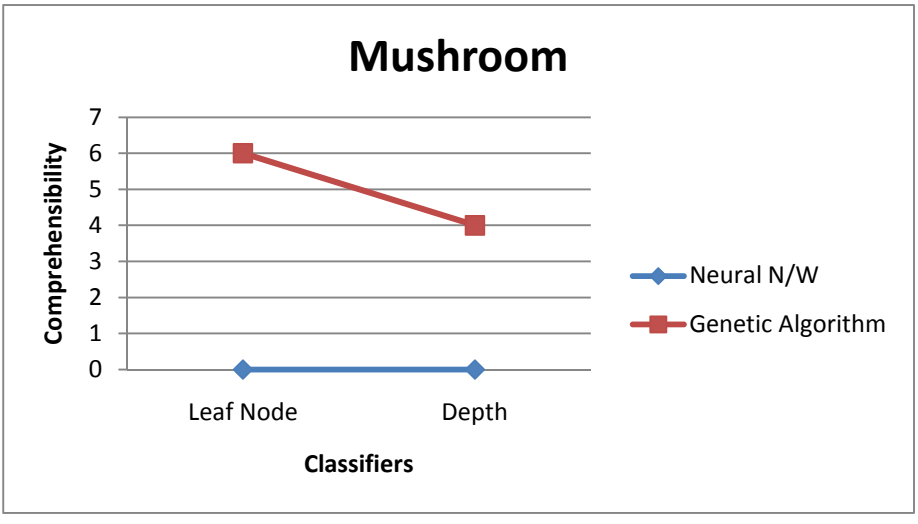


Fig. 3. Comprehensibility on mushroom data set

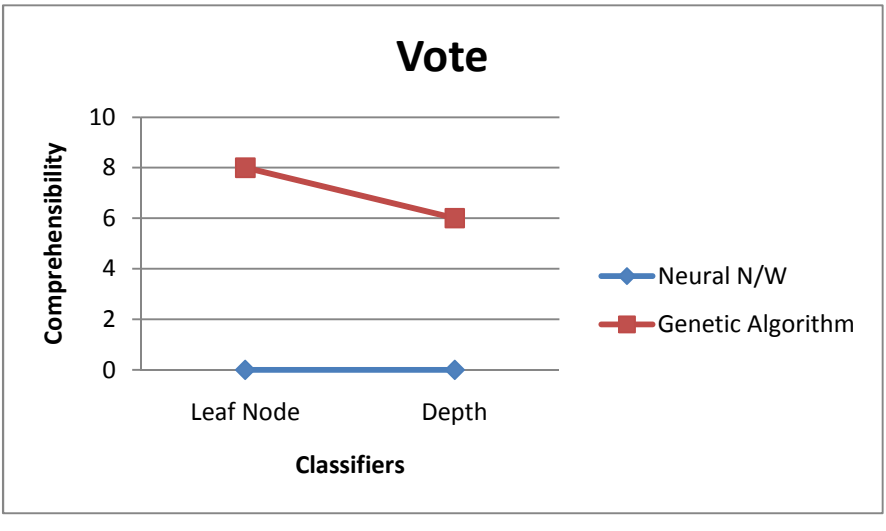


Fig. 4. Comprehensibility on vote data set

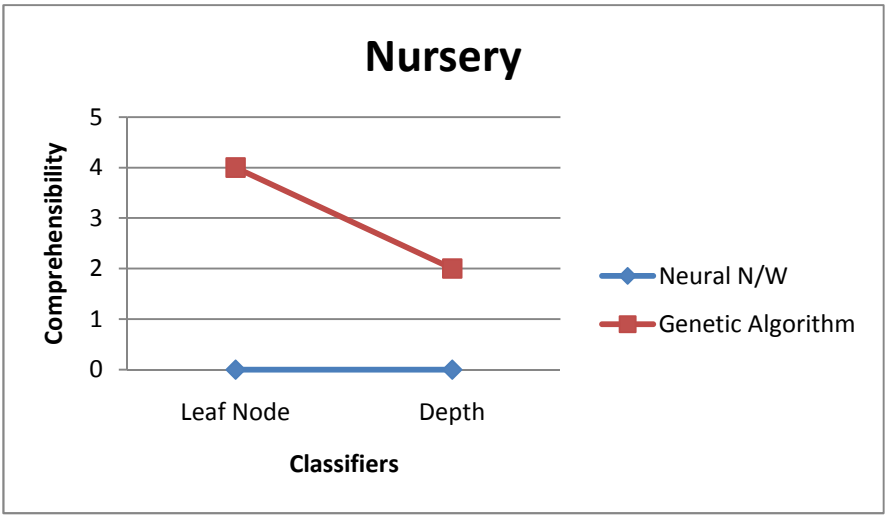


Fig. 5. Comprehensibility on nursery dataset

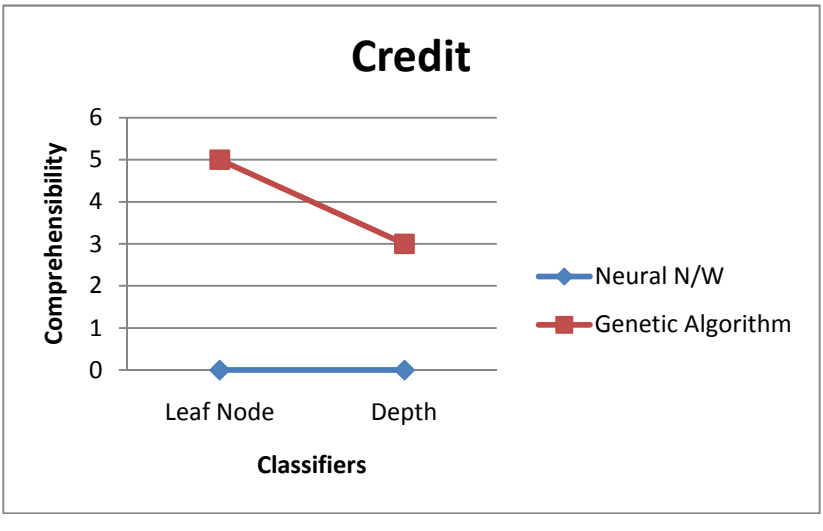


Fig. 6. Comprehensibility on Credit dataset

5 Conclusion

Experimental results demonstrate that evolutionary approach based genetic algorithm classifier is the main choice for organizations when predictive accuracy be the selection criteria as its independence from the datasets domain. Such a classifier is the slowest one due to its chromosomal processing nature. Although ANN based back propagation classifier is also slow but faster than its competitor.

Comprehensibility of evolutionary approach based classifier is excellent as compared to ANN based classifier. So, finally we can conclude that evolutionary approach based classifier should be the first choice of organizations for their decision support systems.

References

1. Fayyad, U.M., Shapiro, G.P., Smyth, P.: The KDD process for extracting useful knowledge from volumes from data. Communication of ACM 39(11), 27–34 (1996)
2. King, R.D., Feng, C., Sutherland, A.: STATLOG. Comparison of classification algorithms on large real-world problems. Applied Artificial Intelligence 9(3), 289–333 (1995)
3. Breslow, L.A., Aha, D.W.: Simplifying decision trees: A survey. Knowledge Engineering Review 12, 1–40 (1997)
4. Brodley, C.E., Utgoff, P.E.: Multivariate versus univariate decision trees. Department of Computer Science, University of Massachusetts, Amherst, MA. Technical Report 92-8 (1992)
5. Brown, D.E., Corruble, V., Pittard, C.L.: A comparison of decision tree classifiers with back propagation neural networks for multimodal classification problems. Pattern Recognition 26, 953–961 (1993)
6. Curram, S.P., Mingers, J.: Neural networks, decision tree induction and discriminant analysis: An empirical comparison. Journal of the Operational Research Society 45, 440–450 (1994)
7. Shavlik, J.W., Mooney, R.J., Towell, G.G.: Symbolic and neural learning algorithms: an empirical comparison. Machine Learning 6, 111–144 (1991)
8. Saroj, Bhardwaj, K.K.: A parallel genetic algorithm approach for automated discovery of censored production rules. In: AIAP 2007 Proceedings of the 25th Conference on Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications, pp. 435–441 (2007)
9. UCI Repository of Machine Learning Databases. Department of Information and Computer Science University of California (1994), `http://www.ics.uci.edu/~mlearn/MLRepositry.html`
10. Han, J., Kamber, M.: Data mining: concepts and techniques: Book (Illustrated), 550 pages (January 2001) ISBN-10: 1558604898, ISBN-13: 9781558604896
11. Al-Ghoneim, K., Kumar, B.V.K.V.: Learning ranks with neural networks. In: Proc. SPIE Applications Science Artificial Neural Networks, vol. 2492, pp. 446–464 (1995)
12. Altman, E.I., Marco, G., Varetto, F.: Corporate distress diagnosis: Comparisons using linear discriminant analysis and neural networks
13. Bharadwaj, K.K., Hewahi, N.M., Brando, M.A.: Adaptive Hierarchical Censored Production Rule-Based System: A Genetic Algorithm Approach. In: Borges, D.L., Kaestner, C.A.A. (eds.) SBIA 1996. LNCS, vol. 1159, pp. 81–90. Springer, Heidelberg (1996)
14. Goldberg, D.E.: Genetic algorithms in search, optimization and machine learning. Addison-Wesley (1989)
15. Deb, K.: Genetic Algorithm in search and optimization: The techniques and Applications. In: Proceeding of Advanced Study Institute on Computational Methods for Engineering Analysis and Design, pp. 12.1—12.25 (1993)

16. Saroj: Genetic ”Algorithm A technique to search complex space”. In: Proceedings of National Seminar on Emerging Dimension in Information Technology, August 10-11, pp. 100–105 (2002)
17. Frietas, A.A.: A survey of evolutionary algorithms for data mining and knowledge discovery, pp. 819–845. Springer, New York (2003)
18. Frietas, A.A.: Data mining and knowledge discovery with evolutionary algorithms, 265 pages (2002) ISBN: 978-3-540-43331-6

Some Concepts of Incomplete Multigranulation Based on Rough Intuitionistic Fuzzy Sets

B.K. Tripathy[1], G.K. Panda[2], and Arnab Mitra[2]

[1] SCSE, VIT University, Vellore- 632 014, India
[2] MITS, Rayagada, Odisha, India
tripathybk@vit.ac.in, gkpmail@sify.com,
mitra.anirban@gmail.com

Abstract. The definition of basic rough sets depends upon a single equivalence relation defined on the universe or several equivalence relations taken one each taken at a time. In the view of granular computing, classical rough set theory is based upon single granulation. The basic rough set model was extended to rough set model based on multi-granulations (MGRS) in [6], where the set approximations are defined by using multi-equivalences on the universe and their properties were investigated. Using the hybridized rough fuzzy set model introduced by Dubois and Prade [2], rough fuzzy set model based on multi-granulation is introduced and studied by Wu and Kou [15]. Topological properties of rough sets introduced by Pawlak in terms of their types were recently studied by Tripathy and Mitra [11]. These were extended to the context of incomplete multi granulation by Tripathy and Raghavan [12]. Recently, the concept of multi-granulations based on rough fuzzy sets by Tripathy and Nagaraju [13]. In a recent paper, Tripathy et al [14] introduced the concept of incomplete multigranulation on rough intuitionistic fuzzy sets (MGRIFS) and studied some of its topological properties. In this paper we continue further by introducing the concept of accuracy measures on MGRIFS and prove some of their properties. Our findings are true for both complete and incomplete intuitionistic fuzzy rough set models based upon multi granulation. The concepts and results established in [13] and [14] open a new direction in the study of multigranulation for further study.

Keywords: Rough Sets, Fuzzy rough sets, Intuitionistic Fuzzy Rough Sets, multi granular fuzzy rough sets and multigranular intuitionistic fuzzy rough sets, accuracy measure.

1 Introduction

Two of the most significant models which have been developed to enhance the modeling capability of basic sets are the notion of fuzzy sets introduced by L.A.Zadeh [16] and the notion of rough set introduced by Pawlak [4, 5]. However, the relative scarcity of equivalence relations led to the development of several extensions of this basic notion. One such extension is the rough sets based upon tolerance relations instead of equivalence relations. These rough sets are sometimes called incomplete

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 683–693.
springerlink.com

rough set models [3]. In the view of granular computing, classical rough set theory is researched by a single granulation. The basic rough set model has been extended to rough set model based on multi-granulations (MGRS) in [6], where the set approximations are defined by using multiple equivalence relations on a universe. Using similar concepts, that is taking multiple tolerance relations instead of multiple equivalence relations; incomplete rough set model based on multi-granulations was introduced in [7]. Several fundamental properties of these types of rough sets have been studied [6, 7, 9].The concept in [6] has been extended to rough fuzzy set model based on multi-granulations by Wu and Kou [15]. Very recently, this concept was generalised to incomplete multi granulation based on rough fuzzy sets [13]. Intuitionistic fuzzy set introduced by Atanassov [1] is a generalised notion of fuzzy sets.

The concept of types of rough sets introduced by Pawlak [5] was further studied by Tripathy et al [10, 11] and was extended to the context of MGRS in [12]. Also, similar kind of study is done by Tripathy and Nagaraju [13] for MGRFS. In this paper we introduce the notion of MGIFRS on complete and incomplete information systems and study similar properties.

2 Definitions and Notations

In this section we put forth several notations and define several concepts, which shall be used by us in presenting our work. The concept of fuzzy sets [16] and rough sets [4, 5] were introduced by Zadeh and Pawlak respectively as models to capture uncertainty information. We avoid presenting these notions here, which are quite familiar by now in the scientific community.

2.1 Rough Fuzzy Sets

Dubois and Prade [2] developed the hybrid models of fuzzy rough sets and rough fuzzy sets. The notion of rough fuzzy sets is defined as follows.

Let (U, R) be an approximation space. Then for any X ϵ F(U), the lower and upper approximations of X with respect to R are given by

$$\underline{R}X(x) = \inf_{y\in[x]_R} X(y) \text{ , for all x } \epsilon \text{ U and} \tag{2.1.1}$$

$$\overline{R}X(x) = \sup_{y\in[x]_R} X(y) \text{ ,for all x U.} \tag{2.1.2}$$

2.2 Multigranular Rough Sets

Qian and Liang defined rough set model based on multi-granulations [10] as follows.

Definition 2.2.1. Let K= (U, **R**) be a knowledge base, **R** be a family of equivalence relations, X ⊆ U and P,Q ϵ **R.**

We define the lower approximation and upper approximation of X in U as

$$\underline{P+Q}(X) = \{x \in U / [x]_P \subseteq X or [x]_Q \subseteq X\} \text{ and} \tag{2.2.1}$$

$$\overline{P+Q}(X) = (P+Q(X^C))^C \tag{2.2.2}$$

2.3 Rough Fuzzy Sets Model Based on Multi-Granulations

The concept of multi-granular rough sets was extended to define fuzzy rough sets based on multigranulation by Wu and Kou [15] as follows:

Definition 2.3.1. Let K = (U, R) be a knowledge base, R be a family of equivalence relations on U and P, Q ∈ R.

For $\forall X \in F(U)$, the lower approximation $\underline{P+Q}(X)$ and upper approximation $\overline{P+Q}(X)$ of X based equivalence relations P, Q are defined as follows:

$$\forall x \in U,\ \underline{P+Q}(X)\ (x) = \inf_{y\in[x]_P} X(y) \vee \inf_{y\in[x]_Q} X(y), \tag{2.3.1}$$

$$\forall x \in U,\ \overline{P+Q}(X)\ (x) = ((P+Q)(X^C))^C(x). \tag{2.3.2}$$

If $\underline{P+Q}(X) = \overline{P+Q}(X)$ then X is called definable, otherwise X is called a fuzzy rough set with respect to multigranulations P and Q. The pair $((\underline{P+Q})(X),(\overline{P+Q})(X))$ is called a MG-fuzzy rough set on multi-granulations P and Q. It has been illustrated in [15] that fuzzy rough sets based on multi-granulations and fuzzy rough sets based on single granulations are different. Out of the many properties of MG-fuzzy rough sets on multi-granulations established in [15] we mention below only those which we are going to use in this paper.

Property 2.3.1. Let K= (U, **R**) be a knowledge base, **R** be a family of equivalence relations. For every X ∈ F(U) and P,Q∈R , the following properties hold true.

$$\underline{P+Q}(X) \subseteq X \subseteq \overline{P+Q}\,(X) \tag{2.3.3}$$

$$\underline{P+Q}\,(X^C) = (\overline{P+Q}\,(X))^C \tag{2.3.4}$$

$$\underline{P+Q}\,(X) = \underline{P}(X) \cup \underline{Q}(X) \tag{2.3.5}$$

$$\overline{P+Q}\,(X) = \underline{P}(X) \cap \underline{Q}(X) \tag{2.3.6}$$

3 Intuitionistic Fuzzy Sets, Rough Intuitionistic Fuzzy Sets and Multigranulation

In case of fuzzy sets the sum of the membership and nonmembership values of an element always sums up to 1. However, in many real life situations this is not true. To handle such situations the notion of intuitionistic fuzzy sets was introduced by Atanassov [1] as an extension of the notion of fuzzy sets. We restrict ourselves here from introducing the definition here and refer to [1].

3.1 Rough Intuitionistic Fuzzy Sets

Extending the notion of rough fuzzy sets introduced by Dubois and Prade, rough intuitionistic fuzzy sets can be defined as follows.

Let (U, R) be an approximation space. Then for any X IF(U), the lower and upper approximations of X with respect to R are given by

$$M(\underline{R}X)(x) = \inf_{y\epsilon[x]R} MX(y) \text{ and } N(\underline{R}X)(x) = \sup_{y\epsilon[x]R} NX(y) \text{for all } x\epsilon U \text{ and} \quad (3.1.1)$$

$$M(\overline{R}X)(x) = \sup_{y\epsilon[x]R} MX(y) \text{ and } N(\overline{R}X)(x) = \inf_{y\epsilon[x]R} NX(y) \text{for all } x\epsilon U. \quad (3.1.2)$$

3.2 Rough Intuitionistic Fuzzy Sets Model Based on Multi-Granulations

In this section we extend the concept of rough fuzzy sets on multigranulation of Wu and Kou [15] to introduce rough intuitionistic fuzzy sets on multigranulation as follows.

Definition 3.2.1. Let K = (U, R) be a knowledge base, R be a family of equivalence relations on U and P, Q ∈ R. $(\underline{P+Q})((\overline{P+Q})(X))$

For ∀X ∈ IF(U) , the lower approximation $(\underline{P+Q})(X)$ and upper approximation $(\overline{P+Q})(X)$ of X based equivalence relations P, Q are defined as follows, ∀x ∈ U

$$M(\underline{P+Q})(X)(x) = \inf_{y\epsilon[x]P} MX(y) \vee \inf_{y\epsilon[x]Q} MX(y), \quad (3.2.2)$$

$$N(\underline{P+Q})(X)(x) = \sup_{y\epsilon[x]P} NX(y) \wedge \sup_{y\epsilon[x]Q} NX(y)$$

$$\forall x\epsilon U, M(\overline{P+Q})(X)(x) = (M(\underline{P+Q})(X^C))^C(x) \quad (3.2.3)$$

$$\text{and} \qquad N(\overline{P+Q})(X)(x) = (N(\underline{P+Q})(X^C))^C(x).$$

$$(\underline{P+Q})(X)(x) = (M(\underline{P+Q})(X)(x), N(\underline{P+Q})(X)(x)), \quad (3.2.4)$$

$$(\overline{P+Q})(X)(x) = (M(\overline{P+Q})(X)(x), N(\overline{P+Q})(X)(x)).$$

If $(\underline{P+Q})(X) = (\overline{P+Q})(X)$ then X is called definable, otherwise X is called an intuitionistic fuzzy rough set with respect to multi-granulations P and Q. The pair $((\underline{P+Q})(X),(\overline{P+Q})(X))$ is called a MG-intuitionistic fuzzy rough set on multi-granulations P and Q.

4 Multi Granulation in Incomplete Information Systems

An information system is an ordered triplet S = (U, AT, f), where U is a finite nonempty set of objects, AT is a finite nonempty set of attributes and f: U → V_a, for any $a \in AT$, where V_a is the domain of any attribute a. In particular, a target information system (IS) is given by S = (U, AT, f, D, g), where D is a finite nonempty set of decision attributes and g_d: $U \rightarrow V_d$ for any $d \in D$, where V_d is the domain of a decision attribute d.

For an IS, any attribute domain V_a may contain the special symbol "*" to indicate that the value of an attribute is unknown. Any domain value different from "*" is called regular.

Definition 4.1. A system in which values of all attributes for all objects from U are regular (known) is called complete and is called incomplete otherwise.

4.1 MGRS in Incomplete Information System

An information system is a 5-tuple S = (U, AT, f, D, g) is called an incomplete target IS if values of some attributes in AT are missing and those of all attributes in D are regular (known), where AT is called the set of conditional attributes and D is the set of decision attributes.

Definition 4.1.1. Let S = (U, A) be an incomplete information system, P ⊆ A an attribute set. We define a binary relation on U as follows

$$\text{SIM(P)} = \{(u, v) \in U \times U \mid \forall a \in P, a(u) = a(v) \text{ or } a(u) = * \text{ or } a(v) = *\}. \quad (4.1.1)$$

If the attributes $P \subseteq AT$ are numerical attributes, we define SIM(P) relation as:

$$\text{SIM(P)} = \{(u, v) \in U \times U \mid \forall a \in P, |a(u) - a(v)| \leq \delta_a \text{ or } a(u) = * \text{ or } a(v) = *\}. \quad (4.1.2)$$

In fact, SIM(P) is a tolerance relation on U. The concept of a tolerance relation has a wide variety of applications in classifications [9,10]. It can be shown that

$$\text{SIM(P)} = \bigcap_{a \in P} \text{SIM}(\{a\}) \quad (4.1.3)$$

Let $S_P(u)$ denote the set $\{v \in U \mid (u,v) \in \text{SIM(P)}\}$. $S_P(u)$ is the maximal set of objects which are possibly indistinguishable by P with u.

Let U/SIM(P) = $\{ S_P(u) \mid u \in U\}$, the classification or the knowledge induced by P. A member $S_P(u)$ from U/SIM(P) will be called a tolerance class or an information granule. It should be noticed that the tolerance classes in U/SIM(P) do not constitute a partition of U in general. They constitute a cover of U, i.e., $S_P(u) \neq \phi$ for every $u \in U$, and $\bigcup_{u \in U} S_P(u) = U$.

Next we define incomplete MGRS on Two Granulation Spaces.

Definition 4.1.2. Let S = (U, AT, f) be an incomplete information system. Let P, Q ⊆ AT be two attribute subsets and X ⊆ U. We define the lower approximation of X and the upper approximation of X in U by the following:

$$\underline{P+Q}\, X = \bigcup \{x \mid S_P(x) \subseteq X \text{ or } S_Q(x) \subseteq X\} \text{ and} \quad (4.1.4)$$

$$\overline{P+Q}\ (X) = (\underline{P+Q}\ (X^c))^c \quad (4.1.5)$$

The ordered pair $((\underline{P+Q})(X),(\overline{P+Q})(X))$ is called a rough set of X with respect to P+Q. The area of uncertainty or boundary region of this rough set is defined by

$$BN_{(P+Q)}(X) = (\overline{P+Q})(X) \setminus (P+Q)(X) \tag{4.1.6}$$

This concept was extended to develop Multi Granulation Rough Fuzzy sets on incomplete information systems in [13] and its topological properties were studied.

4.2 MGRIFS in an Incomplete Information System

In this section we generalise both the MGRFS and MGRS on incomplete information systems to introduce the concept of MGRIFS in incomplete information systems.

Let S = (U, AT, f) be an incomplete target IS and P, Q ⊆ AT two-attribute subsets, and X ∈ IF(U). Then a lower approximation of X in U is defined by

$$M(\underline{P+Q})(X)(x) = \inf_{y\in S_P(x)} MX(y) \vee \inf_{y\in S_Q(x)} MX(y), \forall x \in U; \tag{4.2.1}$$

$$N(\underline{P+Q})(X)(x) = \sup_{y\in S_P(x)} NX(y) \wedge \sup_{y\in S_Q(x)} NX(y), \forall x \in U; \tag{4.2.2}$$

$$M(\overline{P+Q})(X)(x) = (M(\underline{P+Q})(X^C))^C(x), \forall x \in U; \tag{4.2.3}$$

$$N(\overline{P+Q})(X)(x) = (N(\underline{P+Q})(X^C))^C(x), \forall x \in U. \tag{4.2.4}$$

The above definition can be easily extended from two granulations to finite granulations by taking m number of attributes $P_1, P_2, \ldots P_m$ and replacing the operations $\wedge$ by min and $\vee$ by max in (4.2.1) to (4.2.4).

Note 4.2.1. When X is a fuzzy set the above definition reduces to MGRFS as follows.

Here, we have NX(y) = 1 – MX(y) for all $y \in X$. So,

$$\begin{aligned} N(\underline{P+Q})(X)(x) &= \sup_{y\in S_P(x)} NX(y) \wedge \sup_{y\in S_Q(x)} NX(y) \\ &= \sup_{y\in S_P(x)} \{1-MX(y)\} \wedge \sup_{y\in S_Q(x)} \{1-MX(y)\} \\ &= \{1-\inf_{y\in S_P(x)} MX(y)\} \wedge \{1-\inf_{y\in S_Q(x)} MX(y)\} \\ &= 1-\{\inf_{y\in S_P(x)} MX(y) \vee \inf_{y\in S_Q(x)} MX(y)\} \\ &= 1-M(\underline{P+Q})(X)(x). \end{aligned}$$

$$\begin{aligned} N(\overline{P+Q})(X)(x) &= (N(\underline{P+Q})(X^C))^C(x) \\ &= \{1-M(\underline{P+Q})(X^C)\}^C(x) \\ &= 1-\{M(\underline{P+Q})(X^C)\}^C(x) \\ &= 1-M(\overline{P+Q})(X)(x). \end{aligned}$$

Example 4.2.1. Let us consider the following incomplete target IS about an emporium investment project.

Table 1.

Project	Locus	Investment	Population Density	Decision
x_1	common	high	0.88	yes
x_2	Bad	high	*	yes
x_3	Bad	*	0.33	no
x_4	Bad	low	0.40	no
x_5	Bad	low	0.37	no
x_6	Bad	*	0.60	yes
x_7	common	high	0.65	no
x_8	Good	*	0.62	yes

Let K = (U, AT, f, D), where U = $\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}$ and AT = {L, I, P}.

U/SIM (L) = $\{\{ x_1,x_7\},\{x_2,x_3,x_4,x_5,x_6\},\{x_8\}$ and

U/SIM (P) = $\{\{ x_1,x_2\},\{x_1,x_2,x_3,x_4,x_5,x_6,x_7,x_8\}, \{x_2,x_6,x_7,x_8\}\}$.

Suppose, X = $\{(x_1,0.5,0.3), (x_2,0.3,0.4), (x_3,0.3,0.5), (x_4,0.6,0.3), (x_5,0.5,0.4)$ $(x_6,0.8,0.2), (x_7,1,0), (x_8,0.8,0.1)\}$.

$(\underline{L+P})(X)$ = $\{(x_1,0.5,0.3), (x_2,0.3,0.5), (x_3,0.3,0.5), (x_4,0.3,0.5), (x_5,0.3,0.5),$ $(x_6,0.3,0.5), (x_7,0.5,0.3), (x_8,0.8,0.1)\}$

X^C = $\{(x_1,0.3,0.5), (x_2,0.4,0.3), (x_3,0.5,0.3), (x_4,0.3,0.6), (x_5,0.4,0.5),$ $(x_6,0.2,0.8), (x_7,0,1), (x_8,0.1,0.8)\}$

$(\overline{L+P})(X)$ = $\{(x_1,1,0), (x_2,0.8,0.2), (x_3,0.8,0.2), (x_4,0.8,0.2), (x_5,0.8,0.2),$ $(x_6,0.8,0.2), (x_7,1,0), (x_8,0.8,0.1)\}$

5 Accuracy Measure and Topological Properties of MGFRS in Incomplete Information Systems

It has been noted by Pawlak that in the practical applications of rough sets two characteristics are very important. These are the accuracy measure and the topological characterization. The topological characterization of rough sets depends upon the four types of rough sets. Following this approach, we define below four types of MGRIFS in an incomplete information system. Here, we denote by the strict one cut of an intuitionistic fuzzy set X by $X_{<1}$ and it contains all the elements of U which have non-membership value in X strictly less than one.

Note 5.1. It may be noted that in the case of a fuzzy set, this is equivalent to the support set of A, which comprises of elements having positive membership value.

For, by Note 4.2.1 above, for every x in X,

$$M(\underline{P+Q})(X)(x) = 1 - N(\underline{P+Q})(X)(x) \text{ and}$$
$$M(\overline{P+Q})(X)(x) = 1 - N(\overline{P+Q})(X)(x). \text{ So,}$$
$$N(\underline{P+Q})(X)(x) < 1 \Leftrightarrow M(\underline{P+Q})(X)(x) > 0 \text{ and}$$
$$N(\overline{P+Q})(X)(x) < 1 \Leftrightarrow M(\overline{P+Q})(X)(x) > 0.$$

Note 5.2 When the number of attributes is greater than 2 also the above observations hold by parallel logic.

The topological characterization of a MGRIFS was defined in [14] by extending the corresponding notion for MGFRS in [13].We only note here that the definitions given in [13] for two granulations can be extended to finite number of granulations and the properties established there hold true for the general case also.

5.1 Accuracy Measures

In this section we introduce and investigate some accuracy measures and establish their properties. It is well known that the uncertainty of a concept is due to the existence of the borderline region. The accuracy of a concept is inversely proportional to the size of the borderline region. We now introduce the accuracy measure in incomplete MGRIFS.

Definition 5.1.1. Let S = (U, AT, f) be an incomplete IS and $X \in IF(U)$. Let $P = \{P_1, P_2, ...P_m\}$, where $P_i \in AT$ for i = 1, 2,...m. Then the approximation measure of X by P is defined as $\alpha_P(X) = \frac{\left|\left(\left(\sum_{i=1}^{m} \underline{P_i X}\right)\right)_{<1}\right|}{\left|\left(\sum_{i=1}^{m} \overline{P_i} X\right)_{<1}\right|}$, where $X \neq \phi$ and $|X|$ denotes the cardinality of X.

Next, we illustrate the computation of the accuracy measure through an example.

Example 5.1.1.

Let us consider the example 4.3.1 above. Suppose,

$$X=\{(x_1,0,1),(x_2,0.3, 0.6),(x_3,0,0.6),(x_4,0,1),(x_5,0,0.6),(x_6,1,0),(x_7,0,1), (x_8,1,0)\}$$

Then

$$(\underline{L+P})(X) = \{(x_1,0,1), (x_2,0,1),(x_3 ,0,1),(x_4,0,1),(x_5,0,1),(x_6,0,1),(x_7,0,1),(x_8 ,1,0)\}.$$

So that $\left|((L+P)(X))_{<1}\right| = 1.$

$$(\overline{L+P})(X) = \{(x_1,0,1),(x_2,1,0),(x_3,1,0),(x_4,1,0),(x_5,1,0),(x_6,1,0),(x_7,0,1), (x_8 ,1,0)\}.$$

So that $\left|((\overline{L+P})(X))_{<1}\right| = 6.$

Hence, $\alpha_{L+P}(X) = 1/6.$

Similarly, taking X={$(x_1,0,1),(x_2,0,0.6),(x_3,0,0.6),(x_4,0,0.6),(x_5,0,0.6),(x_6,0.8,0.1),(x_7,0,0.9),(x_8,0,1)$},

$(\underline{L+P})(X)$ = {$(x_1,0,1),(x_2,0,0.6),(x_3,0,0.6),(x_4,0,0.6),(x_5,0,0.6),(x_6,0,0.6),(x_7,0,1),(x_8,0,1)$}. Hence, $\left|((L+P)(X))_{<1}\right| = 5.$

Also, $(\overline{L+P})(X)$ = {$(x_1,0,0.9)$, $(x_2,0.8,0.1),(x_3,0.8,0.1),(x_4,0.8,0.1),(x_5,0.8,0.1),(x_6,0.8,0.1),(x_7,0,0.9),(x_8,0,1)$}. ,

So that $\left|\left((\overline{L+P})(X)\right)_{<1}\right| = 7.$ Hence, $\alpha_{L+P}(X) = 5/7.$

Now, we shall establish some properties of the accuracy measure. We shall use the following result in establishing the next property of accuracy measure.

Property 5.1.1. Let S = (U, AT, f) be an incomplete IS and X be in IF(U). Then for $P_1, P_2, ...P_m \in AT$,

$$\underline{\sum_{i=1}^{m} P_i X} = \bigcup_{i=1}^{m} \underline{P_i} X \quad \text{and} \tag{5.1.1}$$

$$\overline{\sum_{i=1}^{m} P_i X} = \bigcap_{i=1}^{m} \overline{P_i} X. \tag{5.1.2}$$

Theorem 5.1.1. Let S = (U, AT, f) be an incomplete IS and $X \in IF(U), P = \{P_1, P_2, ...P_m\}, where\ P_i \in AT\ for\ i = 1,2,...m.$ P' be a subset of P. Then

$$\alpha_P(X) \geq \alpha_{P'}(X) \geq \alpha_{P_j}(X), for\ each\ P_j \in P'.$$

Proof. Since $P' \subseteq P$, we have $\bigcup_{i=1}^{m} \underline{P_i}(X) \supseteq \bigcup_{P_j \in P'} \underline{P_i}(X)$ *and* $\bigcap_{i=1}^{m} \overline{P_i}(X) \subseteq \bigcap_{P_j \in P'} \overline{P_j}(X).$

So, $\left(\bigcup_{i=1}^{m} \underline{P_i}(X)\right)_{<1} \supseteq \left(\bigcup_{P_j \in P'} \underline{P_i}(X)\right)_{<1}$ *and* $\left(\bigcap_{i=1}^{m} \overline{P_i}(X)\right)_{<1} \subseteq \left(\bigcap_{P_j \in P'} \overline{P_j}(X)\right)_{<1}.$

Hence, $\left|\left(\bigcup_{i=1}^{m} \underline{P_i}(X)\right)_{<1}\right| \geq \left|\left(\bigcup_{P_j \in P'} \underline{P_i}(X)\right)_{<1}\right|$ *and* $\left|\left(\bigcap_{i=1}^{m} \overline{P_i}(X)\right)_{<1}\right| \leq \left|\left(\bigcap_{P_j \in P'} \overline{P_j}(X)\right)_{<1}\right|.$

Thus $\alpha_{P}(X) = \dfrac{\left|\left(\underline{\sum_{i=1}^{m} P_i X}\right)_{<1}\right|}{\left|\left(\overline{\sum_{i=1}^{m} P_i X}\right)_{<1}\right|} = \dfrac{\left|\left(\bigcup_{i=1}^{m} \underline{P_i}(X)\right)_{<1}\right|}{\left|\left(\bigcap_{i=1}^{m} \overline{P_i}(X)\right)_{<1}\right|} \geq \dfrac{\left|\left(\bigcup_{P_j \in P'} \underline{P_i}(X)\right)_{<1}\right|}{\left|\left(\bigcap_{P_j \in P'} \overline{P_j}(X)\right)_{<1}\right|} = \dfrac{\left|\underline{\sum_{P_i \in P'} P_i X}\right|}{\left|\overline{\sum_{P_i \in P'} P_i X}\right|} = \alpha_{P'}(X).$

As $\{P_j\} \subseteq P'\ for\ each\ P_j \in P', we\ get \alpha_{P'}(X) \geq \alpha_{P_j}(X).$

5.1.1 Some Deductions

We can derive an accuracy measure for MGRFS from definition 5.1 as follows.

Definition 5.1.1.1. Let S = (U, AT, f) be an incomplete IS and $X \subseteq F(U).$ Let $P = \{P_1, P_2, ...P_m\}$, where $P_i \in AT$ for i = 1, 2, ...m. Then the approximation measure of X by P is

defined as $\alpha_P(X)=\frac{\left|\left(\underline{(\sum_{i=1}^{m} P_i X)}\right)_{>0}\right|}{\left|\left(\overline{(\sum_{i=1}^{m} P_i X)}\right)_{>0}\right|}$, where $X \neq \phi$ and $|X|$ denotes the cardinality of X. Here, for any fuzzy set X $(X)_{>0}$ represents the set of elements in whose membership value is positive or we can say it is the support set of X.

Properties similar to that in Theorem 5.2.1 can obtained. When X is a crisp set,

$$\left(\underline{(\sum_{i=1}^{m} P_i X)}\right)_{>0} \text{ and } \left(\overline{(\sum_{i=1}^{m} P_i X)}\right)_{>0} \text{ reduce to } \left(\underline{(\sum_{i=1}^{m} P_i X)}\right)_{>0} \text{ and } \left(\overline{(\sum_{i=1}^{m} P_i X)}\right)_{>0}$$

respectively and $\alpha_P(X)$ reduces to its definition given in [15].

6 Conclusions

In this paper we extended the definition of the concept of multigranular intuitionistic fuzzy rough sets in incomplete information systems introduced in [14] and introduced the concepts of accuracy measures on these systems. We also established some properties of these measures. The topological properties of MGIFRSs established in [14] can be extended to this context. So, we have introduced the two concepts needed for application of such generalised models in real life problems. However, the applications of these concepts are needed to be explored further. These results can be used in approximation of classifications and rule induction.

References

[1] Atanassov, K.T.: Intuitionistic Fuzzy Sets. Fuzzy Sets and Systems 20, 87–96 (1986)
[2] Dubois, D., Prade, H.: Rough fuzzy sets model. J. of General Systems 46(1), 191–208 (1990)
[3] Kryszkiewicz, K.: Rough set approach to incomplete information systems. J. Information Sciences 112, 39–49 (1998)
[4] Pawlak, Z.: Rough sets. J. of Computer and Information Sciences 11, 341–356 (1982)
[5] Pawlak, Z.: Rough sets: Theoretical aspects of reasoning about data. Kluwer academic publishers (1991)
[6] Qian, Y.H., Liang, J.Y.: Rough set method based on Multi-granulations. In: Proc of the 5th IEEE Conference on Cognitive Informatics, vol. 1, pp. 297–304 (2006)
[7] Qian, Y.H., Liang, J.Y., Dang, C.Y.: MGRS in Incomplete Information Systems. In: IEEE Conf. on Granular Computing, pp. 163–168 (2007)
[8] Qian, Y.H., Liang, J.Y., Dang, C.Y.: Incomplete Multigranulation Rough set. IEEE Transactions on Systems, Man and Cybernetics-Part A: Systems and Humans 40(2), 420–431 (2010)
[9] Tripathy, B.K.: On Approximation of classifications, rough equalities and rough equivalences. In: Rough Set Theory: A True Landmark in Data Analysis. SCI, vol. 174, pp. 85–136. Springer (2009)
[10] Tripathy, B.K.: Rough Sets on Fuzzy Approximation Spaces and Intuitionistic Fuzzy Approximation Spaces. In: Rough Set Theory: A True Landmark in Data Analysis. SCI, vol. 174, pp. 3–44. Springer (2009)

[11] Tripathy, B.K., Mitra, A.: Topological Properties of Rough Sets and their Applications. J. of Granular Computing, Rough Sets and Intelligent Systems (Switzerland) 1(4), 355–369 (2010)
[12] Tripathy, B.K., Raghavan, R.: On Some Topological Properties of Multigranular Rough Sets. J. of Advances in Applied Science Research 2(3), 536–543 (2011)
[13] Tripathy, B.K., Nagaraju, M.: Topological Properties of Incomplete Multigranulation Based on Rough Fuzzy Sets. In: Proc. of ObCom 2011, pp. 94–103 (2011)
[14] Tripathy, B.K., Panda, G.K., Mitra, A.: Incomplete Multigranulation Based on Rough Intuitionistic Fuzzy Sets. J. of UNIASCIT 2(1), 118–124 (2011)
[15] Wu, M., Kou, G.: Fuzzy Rough Set Model on Multi-Granulations. In: Proc of the 2nd Int. Conf. on Computer Engineering and Technology, vol. 2, pp. 72–75 (2010)
[16] Zadeh, L.: Fuzzy Sets. J. Information and Control 8(11), 338–353 (1965)

Data Mining Model Building as a Support for Decision Making in Production Management

Pavol Tanuska, Pavel Vazan, Michal Kebisek, Oliver Moravcik, and Peter Schreiber

Institute of Applied Informatics, Automation and Mathematics, The Faculty of Materials Science and Technology Slovak University of Technology, Trnava, Slovakia
pavol.tanuska@stuba.sk, pavel.vazan@stuba.sk, michal.kebisek@stuba.sk, oliver.moravcik@stuba.sk, peter.schreiber@stuba.sk

Abstract. The paper gives the next stages of the project that is oriented on the use of data mining techniques and knowledge discoveries from production systems through them. They have been used in the management of these systems. Production data was obtained in previous stages of project. This production data are stored in data warehouse that was proposed and developed by authors. Data mining model has been created by using specific methods and selected techniques for defined problems of production system management. The main focus of our article is the proposal of data mining model.

Keywords: Multidimensional Model, Data Warehouse, Data Mining, Data Mining Model, Production Goals.

Related Works

The paper presents the proposal of solution procedure of the research project that is oriented to the knowledge discovery from production system databases. The pa-per gives the first pilot results of our solution. The authors published the first stages of solution procedure in [1]. The next stages were published by authors of the research team [2].

1 Introduction

Today we are witnesses of a continuous development of information technology. It grows exponentially amount of generated and stored data that is available. This often does not bring the better knowing but mainly contributes to the disorientation and inability to decide objectively. The excess of data which shall allow to the responsible operator doing of qualified decisions, does not lead to understanding of the situation, but often to disorientation and a time stress. Available data still does not mean anything to know. Information system filled with data even though has its meaning and form the basis of the "memory" of the organization. The areas having the ability to generate and store large quantities of data have not only areas such as marketing,

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 695–701.
springerlink.com

government, or medicine, but this trend begins to observe in industrial areas increasingly. At the designing of new production lines and equipment is calculated with the possibility of collecting and storing business data from the manufacturing process. Older equipment and lines are often adjusted in order to add this option. It is not necessary to collect such data only, but it is necessary to work with these data properly. Analysts need to obtain information to be able to model objects, anticipate trends and to enable the responsible managers to make appropriate decisions. Classic methods based on SQL, do not enough for it already. For that reason it is the effort to put into this process the methods of knowledge discovery, whose task is to extract a new, valid and potentially useful knowledge from large volumes of available data [3]. In our article we will discuss the possibility of the application process of knowledge discovering in databases in the industrial area.

2 Project Objectives

Production management must ensure the achievement of different production goals in a given timeframe. These objectives are often conflicting and their achievement depends on many factors. Many dependencies are so far very little explored, for example relationship of capacity utilization and value of flow time, depending on the size of the production batch [2]. The problem of minimizing variable costs depending on the necessary operating supplies possession, alternatively with the possibility of increasing the value added percentage parameter (metric according to Lean Production). There are also few issues that are very little investigated, like the impact of priority rules for allocating the operations on the production goals. These problems and dependencies would be possible to solve by using data mining methods [4]. The process of knowledge discoveries from management of the production systems can be applied for solution of many problems. Here belong:

- identification of production parameters influence on a production process,
- identification of breakdowns in production process,
- deviations (divergences) detection from plan during the production,
- failure states detection of production equipments,
- production process optimisation,
- workstations layout optimisation,
- failures prediction in production process, etc.

2.1 Conceptual Model of the Project

The whole process of obtaining, storing and pre-processing data was divided into several stages. These stages are shown in Fig. 2.1. The authors expected that suggested procedure would discover new knowledge. Then the new control strategies will be defined on the base of the new knowledge. The first three stages of conceptual model were solved in [1]. The next stages of the proposed conceptual model will be presented in this paper.

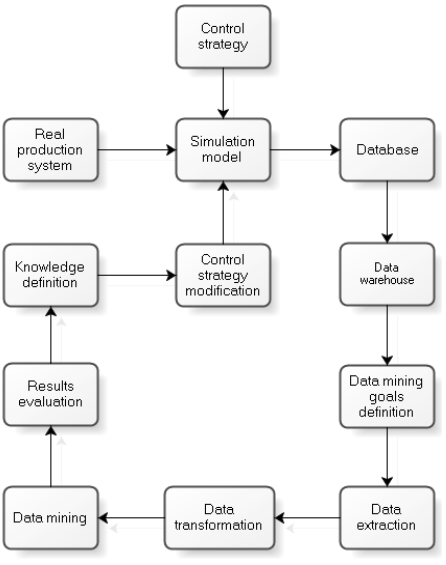


Fig. 2.1. The conceptual model of the project

2.2 Proposal of Multidimensional Model

The data warehouse was designed and built by using Oracle Data warehouse Builder 11g. Multidimensional model of data warehouse in Fig. 2.2 is designed as a star

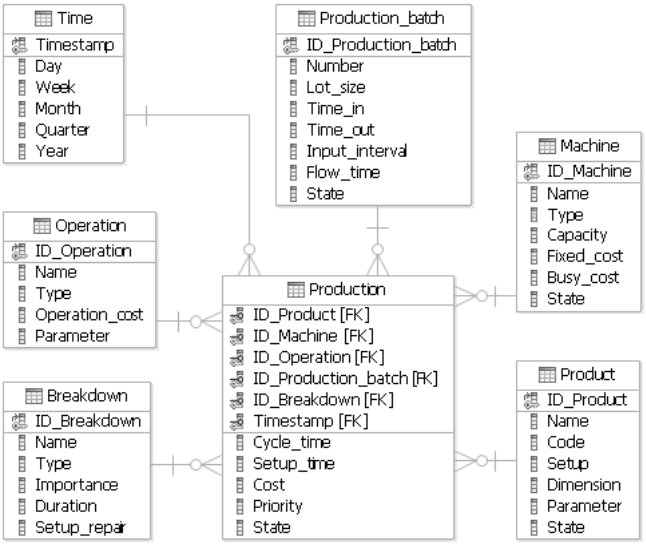


Fig. 2.2. Proposal of multidimensional model

scheme and uses the fact table and six dimensional tables. Some data mining tools allow direct connection to the relational databases. Data obtained from relationship databases have to be cleaned and adjusted directly to the data mining tools. Therefore data analyses are complicated and lengthy. The data mining tool collects the preprocessed and fully transformed data from the proposed data warehouse. The whole process of data mining was facilitated and accelerated by creating the data warehouse.

3 Design of Data Mining Model

The following steps were executed during the proposal process of the data mining model. The first step was to determine of data mining process goals. Following these goals authors made a choice of relevant data set from proposed and developed data warehouse. The next step was the creation of mentioned data mining model.

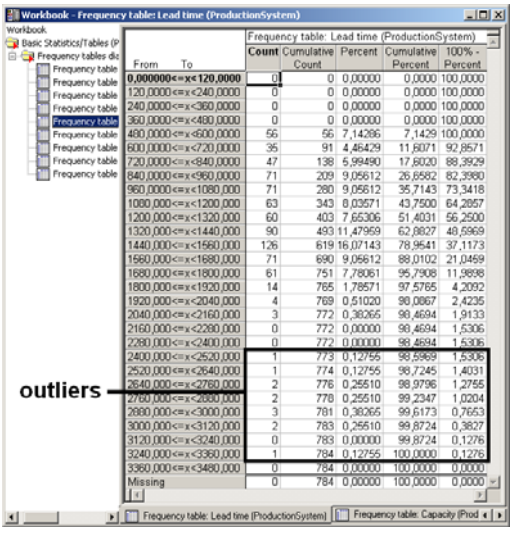
3.1 Data Mining Goals

We have defined goals that we try to obtain by using knowledge discovery from databases process in the first stage [2]:

- the influence analysis of production process parameters for production equipment utilization,
- the influence analysis of production process parameters for flow time of production batches,
- the influence analysis of production process parameters for number of finished parts.

3.2 Preprocessing and Transformation

Authors came out from modified and transformed data warehouse data set during the data selection. The data warehouse was proposed and developed in previous steps of this process. In the proposed model we utilised possibilities that model provides: the data was tested on missing values, on values distribution, identification of extreme data values was executed and the final step was data adjusting and transformation [5]. This adjusted and transformed data set served as an input set into next steps of data mining process. We have adjusted and transformed the input data set. The next logical step has been investigation whether data set does not contain data which values are markedly different from others data so-called outliers (Fig. 3.1). It would be necessary to discover whether data set contains this kind of data and also the reason of their existence. It could be a random data and their extreme values could be caused by an error at the data recording. But it could be a significant data too. Therefore it has been important to identify origin of these values correctly and decide whether these values will be included in used data set or will be removed [6]. We have used the possibility of creation “Frequency” table for identification of markedly different data. Moreover the data set has been examined whether it does not contain missing data or kind of interference in some parameters [7]. We have used transformed and adapted input data set in the next process steps [2].



From To	Count	Cumulative Count	Percent	Cumulative Percent	100% - Percent
0,000000<=x<120,0000	0	0	0,00000	0,0000	100,0000
120,0000<=x<240,0000	0	0	0,00000	0,0000	100,0000
240,0000<=x<360,0000	0	0	0,00000	0,0000	100,0000
360,0000<=x<480,0000	0	0	0,00000	0,0000	100,0000
480,0000<=x<600,0000	56	56	7,14286	7,1429	100,0000
600,0000<=x<720,0000	35	91	4,46429	11,6071	92,8571
720,0000<=x<840,0000	47	138	5,99490	17,6020	88,3929
840,0000<=x<960,0000	71	209	9,05612	26,6582	82,3980
960,0000<=x<1080,000	71	280	9,05612	35,7143	73,3418
1080,000<=x<1200,000	63	343	8,03571	43,7500	64,2857
1200,000<=x<1320,000	60	403	7,65306	51,4031	56,2500
1320,000<=x<1440,000	90	493	11,47959	62,8827	48,5969
1440,000<=x<1560,000	126	619	16,07143	78,9541	37,1173
1560,000<=x<1680,000	71	690	9,05612	88,0102	21,0459
1680,000<=x<1800,000	61	751	7,78061	95,7908	11,9898
1800,000<=x<1920,000	14	765	1,78571	97,5765	4,2092
1920,000<=x<2040,000	4	769	0,51020	98,0867	2,4235
2040,000<=x<2160,000	3	772	0,38265	98,4694	1,9133
2160,000<=x<2280,000	0	772	0,00000	98,4694	1,5306
2280,000<=x<2400,000	0	772	0,00000	98,4694	1,5306
2400,000<=x<2520,000	1	773	0,12755	98,5969	1,5306
2520,000<=x<2640,000	1	774	0,12755	98,7245	1,4031
2640,000<=x<2760,000	2	776	0,25510	98,9796	1,2755
2760,000<=x<2880,000	2	778	0,25510	99,2347	1,0204
2880,000<=x<3000,000	3	781	0,38265	99,6173	0,7653
3000,000<=x<3120,000	2	783	0,25510	99,8724	0,3827
3120,000<=x<3240,000	0	783	0,00000	99,8724	0,1276
3240,000<=x<3360,000	1	784	0,12755	100,0000	0,1276
3360,000<=x<3480,000	0	784	0,00000	100,0000	0,0000
Missing	0	784	0,00000	100,0000	0,0000

Fig. 3.1. Outliers identification

The STATISTICA Data Miner provides various interactive tools for data analysing. These tools were used for first view creation into analysed data. We detected decomposition of values for individual analysed attributes, e.g. min and max values, values variance and so on. Having finished these partial fractions we can continue with next steps in the knowledge discovery from databases. The "Feature selection and Variable screening" models were used to detect which data can have influence into spotted parameters. These models identified and presented dependencies between individual analysed parameters (Fig. 3.2).

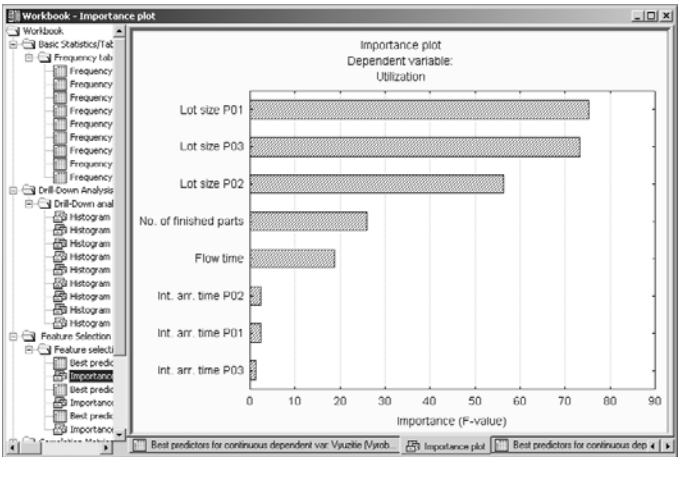


Fig. 3.2. Importance table of relevant parameters

We identified the parameters with the most significant influence upon the investigate goals on the base of obtained results. We proved that the lot sizes of batches have the most significant influence on the three main investigated goals (flow time, number of finished parts and capacity utilisation). The next analyse will be focused on these parameters.

3.3 Designed Data Mining Model

The next stage of data mining model proposal was the selection of methods and techniques of data mining [8]. There were involved the following methods and techniques of data mining into the proposed model (Fig. 3.3):

- STATISTICA Automated Neural Network Regression – Custom Neural Network,
- STATISTICA Automated Neural Network Regression – Automated Network Search,
- Generalized K-Means Cluster Analysis,
- MARSplines,
- SVM (Support Vector Machines).

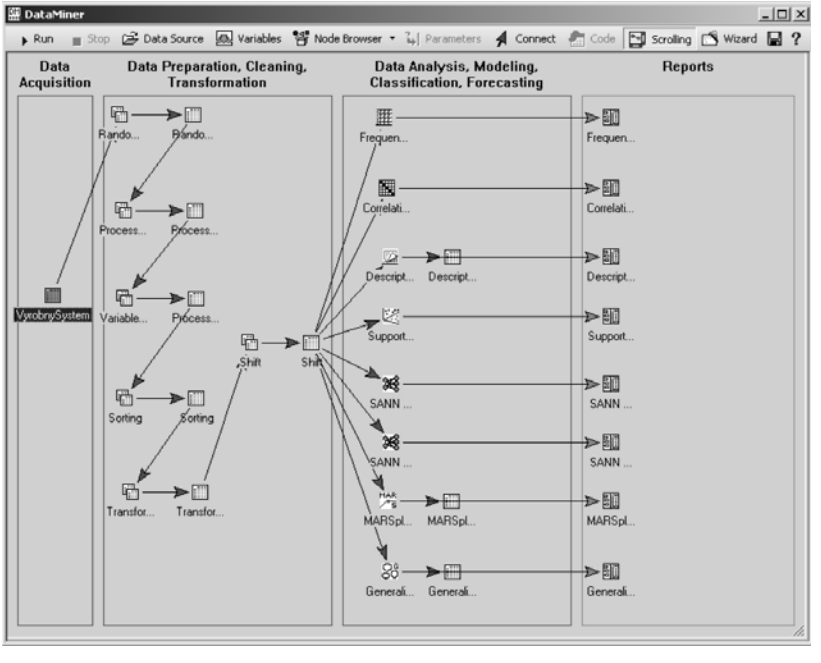


Fig. 3.3. Proposed data mining model

The advantage of usage of Data Miner tool is also the possibility to realize the next modifications of used methods. For the definition of the individual components of the model the software tool STATISTICA Data Miner uses STATISTICA Visual Basic, programming language, which is the modification of programming language

Microsoft Visual Basic. We applied such possibility for particular methods, where it was not possible to setup specific behaviour by the graphical interface but it was necessary to perform the modification of source code directly. The modified method was saved to the user's methods. This method was later used for the next analysed problems.

4 Conclusions

The process of Knowledge Discovery in Databases needs appropriately adjusted data from which obtains valid, recent, till undisclosed, potential usable and well comprehensible knowledge. For this purpose it is necessary to collect appropriate data from the production system. Thus collected data and stored in data warehouses can be used as input data set in the process of Knowledge Discovery in Databases.

The next stages of project will be implementation of knowledge discovery in the production systems management area can be used to achieve a better understanding of a production system. It can be also used to gain new and interesting knowledge for predicting future behaviour of the production system. The new discovered knowledge will help managers in their decision making.

Acknowledgements. This contribution was written with a financial support VEGA agency in the frame of the project 1/0214/11 „The data mining usage in manufacturing systems control".

References

[1] Vazan, P., Kebisek, M., Tanuska, P., Jurovata, D.: The data warehouse suggestion for production system. In: Annals of DAAAM and Proceedings of DAAAM Symposium, Vienna, Austria, vol. 22(1), pp. 0017–0018 (2011) ISBN 978-3-901509-83-4

[2] Vazan, P., Tanuska, P., Kebisek, M.: The data mining usage in production system management. World Academy of Science, Engineering and Technology 7(77), 1304–1308 (2011) ISSN 2010-376X

[3] Giudici, P., Figini, S.: Applied Data Mining for Business and Industry, 2nd edn. John Wiley & Sons Ltd, Cornwall (2009)

[4] Larose, D.: Data Mining Methods and Models. John Wiley & Sons Ltd, New Jersey (2006)

[5] Vrabel, R.: Boundary layer phenomenon for three-point boundary value problem for the nonlinear singularly perturbed systems. Kybernetika 47(4), 644–652 (2011) ISSN 0023-5954

[6] Good, P.: A Practitioner's Guide to Resampling for Data Analysis, Data Mining, and Modeling. Chapman & Hall, London (2011)

[7] Kargupta, H., Han, J.: Next Generation of Data Mining. CRC Press, Boca Raton (2008)

[8] Vercellis, C.: Business Intelligence: Data Mining and Optimization for Decision Making. John Wiley & Sons Ltd, Cornwall (2009)

Multi-Objective Zonal Reactive Power Market Clearing Model for Improving Voltage Stability in Electricity Markets Using HFMOEA

Ashish Saini* and Amit Saraswat**

Department of Electrical Engineering, Faculty of Engineering,
Dayalbagh Educational Institute, Agra 282110, Uttar Pradesh, India
{ashish7119,amitsaras}@gmail.com

Abstract. This paper presents a development of a new multi-objective zonal reactive power market clearing (ZRPMC-VS) model for improving voltage stability of power system. In proposed multi-objective ZRPMC-VS model, two objective functions such as total payment function (TPF) for rective power support from generators and syncronus condensers and voltage stability enhancement index (VSEI) are optimized symultanously by satisfying various power system constraints using hybrid fuzzy multi-objective evolutionary algorithm (HFMOEA). The results obtained using HFMOEA are comapared with a well known NSGA-II solution technique. This analysis helps the independent system operators (ISO) to take better decisions in clearing the reactive power market in competetive market environment.

Keywords: Zonal reactive power market, reactive power market clearing prices, hybrid fuzzy multi-objective evolutionary algorithms.

1 Introduction

In competitive electricity markets engineers view the reactive power management problem from two angles - technical and as well as economical [1-2]. A market model process to manage reactive services by independent transmission operators are presented in [3] and uses a piece-wise linear representation of the capability curve of each generator for computing reactive power cost curves. Zhong et al. [4-5] developed a competitive market for reactive power and raised many important issues of reactive power management. In [4], a two-step approach for reactive power procurement is proposed. This approach is extended in [5] by developing a uniform price auction model for competitive reactive power markets to determine the prices for different components of reactive power services namely: availability, operation

* Ashish Saini, Associate Professor.
** Amit Saraswat, Research Scholar in the Department of Electrical Engineering, Dayalbagh Educational Institute, Agra-282110, Uttar Pradesh, India. (Corresponding author phone: +91-562-2801224; fax: +91-562-2801226; e-mail: ashish7119@gmail.com and amitsaras@gmail.com).

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 703–714.
springerlink.com

and opportunity. Market clearing was achieved by simultaneously considering minimization of payment, total system losses, and deviations from contracted transactions using compromise programming approach. In [6], a localized or zonal reactive power market is proposed using the concept of voltage control areas (VCAs) where the reactive power is settled by calculating the zonal uniform market clearing prices (ZUMCPs). In all these papers, the optimization problems are tackled in single objective optimization framework.

In recent years, all real world optimization problems are being tried to be formulated in multi-objective optimization framework, in which multiple objective functions are optimized simultaneously. Multi-objective optimization problems (MOPs) with are non-commensurable and conflicting objective functions give rise to a set of optimal solutions, instead of one optimal solution, called as pareto-optimal solutions [7]. Several evolutionary multi-objective solution techniques such as Strength Pareto Evolutionary Algorithm (SPEA) [8], fuzzy adaptive particle swarm optimization (FAPSO) [9], a seekers optimization algorithm (SOA) [10], a modified non-dominated sorting genetic algorithm (MNSGA-II) [11] are applied to reactive power optimization problems such as optimal reactive power dispatch ORPD [8-11], and RPMC [12].

In present paper, the multi-objective zonal reactive power market clearing considering voltage stability (ZRPMC-VS) Model is developed as a mixed integer nonlinear programing (MINLP) multi-objective optimization problem which includes two objectives such as total payment function (TPF) for rective power support from generators and syncronus condensers [6] and voltage stability enhancement index (VSEI) [13-14]. A hybrid fuzzy multi-objective evolutionary algorithm (HFMOEA) proposed in reference [15] is applied for solving this multi-objective ZRPMC-VS model and tested on IEEE 24 bus reliability test system. The result obtained in proposed multi-objective ZRPMC-VS using HFMOEA are compared with a well known multi-objective solution technique such as NSGA-II [16].

2 Proposed Multi-Objective ZRPMC-VS Model

In proposed Multi-Objective ZRPMC-VS model, two objective functions such as Total Payment Function (TPF) and Voltage Stability Enhancement Index (VSEI) are optimized simultaneously. The first objective i.e. TPF is a total payment received for providing reactive power services from all generators and synchronous condensers participated in electricity market and being calculated in zonal bases as described in [6]. The TPF for zonal reactive supports may be formulated as follows:

$$F_1 = TPF = \sum_{i \in N_{PV}} \left(\rho_0 . W_{0,i}\right) + \sum_{i \in N_{PV}, N} \begin{pmatrix} -\rho_{1N} . W_{1,iN} . Q_{G1,iN} + \rho_{2N} . W_{2,iN} . Q_{G2,iN} \\ +\rho_{2N} . W_{3,iN} . Q_{GA,iN} + \frac{1}{2} \rho_{3N} . W_{3,iN} . Q_{G3,iN}^2 \end{pmatrix} \quad (1)$$

Reactive power output from i^{th} provider is classified into three components $Q_{G1,i}$, $Q_{G2,i}$ or $Q_{G3,i}$ that represent the regions $\left(Q_{Gmin,i}, 0\right)$, $\left(Q_{Gbase,i}, Q_{GA,i}\right)$ and $\left(Q_{GA,i}, Q_{GB,i}\right)$, respectively. Accordingly, only one of the binary variables W_1, W_2 and W_3 can be

selected. In (1), ρ_0 is the uniform availability price for whole system and ρ_{1N} and ρ_{2N} are the uniform cost of loss prices in Zone-*N*, whereas ρ_{3N} is the uniform opportunity price in Zone-*N*. If a provider is selected, W_0 will be one, and it will receive the availability price, irrespective of its reactive power output.

In this Multi-Objective ZRPMC-VS model, a voltage stability enhancement index (VSEI) also known as *L-index* [13] is consider as second objective. It is a static voltage stability measure of power system computed based on normal load flow solution and its value is computed for each load bus in the system [13-14]. The formulation of VSEI is as follows:

$$F_2 = VSEI = L-index = \max\left\{ L_j = \left| 1 - \sum_{i=1}^{N_{PV}} F_{ji} \frac{V_i}{V_j} \right|, j \in N_{PQ} \right\} \tag{2}$$

All the terms within the sigma of (2) are complex quantities. The values F_{ji} are obtained from the Y-bus matrix as below in (3).

$$\begin{bmatrix} I_G \\ I_L \end{bmatrix} = \begin{bmatrix} Y_{GG} & Y_{GL} \\ Y_{LG} & Y_{LL} \end{bmatrix} \begin{bmatrix} V_G \\ V_L \end{bmatrix} \tag{3}$$

Where $[I_G], [I_L]$ and $[V_G], [V_L]$ represents the complex currents and bus voltages respectively; whereas $[Y_{GG}]$, $[Y_{GL}]$, $[Y_{LG}]$ and $[Y_{LL}]$ are corresponding portions of network Y-bus matrix.

Rearranging (3) we obtain

$$\begin{bmatrix} V_L \\ I_G \end{bmatrix} = \begin{bmatrix} Z_{LL} & F_{LG} \\ R_{GL} & Y_{GG} \end{bmatrix} \begin{bmatrix} I_L \\ V_G \end{bmatrix} \tag{4}$$

$$\text{Where } F_{LG} = -[Y_{LL}]^{-1}[Y_{LG}] \tag{5}$$

The value of *L-index* or VSEI lies between 0 and 1 [14]. A VSEI value less than 1 (voltage collapse point) and close to 0 (no load point) indicates a system state i.e. system voltage stability margin.

Both above objective functions (1 and 2) are optimized while maintain the various system equality and inequality constraints described as follows:

Load Flow Equality Constraints:

$$0 = P_{G,i} - P_{D,i} - V_i \sum_{j \in N_i} V_j \left(G_{ij} \cos\theta_{ij} + B_{ij} \sin\theta_{ij} \right); i \in N_B \tag{6}$$

$$0 = Q_{G,i} - Q_{D,i} - V_i \sum_{j \in N_i} V_j \left(G_{ij} \sin\theta_{ij} + B_{ij} \cos\theta_{ij} \right); i \in N_{PQ} \tag{7}$$

Reactive Power Relational Constraints and Limits

$$Q_{Gi} = Q_{G1,i} + Q_{G2,i} + Q_{G3,i} \tag{8}$$

$$W_{1,i}.Q_{Gmin,i} \le Q_{G1,i} \le W_{1,i}.Q_{Gbase,i} \tag{9}$$

$$W_{2,i}.Q_{Gbase,i} \le Q_{G2,i} \le W_{2i}.Q_{GA,i} \tag{10}$$

$$W_{3,i}.Q_{GA,i} \le Q_{G3,i} \le W_{3,i}.Q_{GB,i} \tag{11}$$

$$W_{1,i} + W_{2,i} + W_{3,i} \le 1 \tag{12}$$

Determining the Market Prices

$$W_{0,i} = W_{1,i} + W_{2,i} + W_{3,i}, \quad \forall i \in N_{PV} \tag{13}$$

$$W_{0,i}.a_{0,i} \le \rho_0, \quad \forall i \in N_{PV} \tag{14}$$

$$W_{1,i}.m_{1,i} \le \rho_{1N}, \quad \forall i \in \text{Zone} - N \tag{15}$$

$$\left(W_{2,i} + W_{3,i}\right).m_{2,i} \le \rho_{2N}, \quad \forall i \in \text{Zone} - N \tag{16}$$

$$W_{3,i}.m_{3,i} \le \rho_{3N}, \quad \forall i \in \text{Zone} - N \tag{17}$$

Reactive Power Generation Limits

$$Q_{G\min,i} \le Q_{G,i} \le Q_{G\max,i}; \quad i \in N_{PV} \tag{18}$$

$$Q_{C\min,i} \le Q_{C,i} \le Q_{C\max,i}; \quad i \in N_C \tag{19}$$

Reactive Power Capability Limits of Generators

$$Q^{\text{limit}}{}_{G,i} \le \begin{cases} \sqrt{\left(V_{t,i}I_{a,i}\right)^2 - P_{G,i}^2} & \text{if } P_{G,i} \ge P_{GR,i} \\ \sqrt{\left(\dfrac{V_{t,i}E_{af,i}}{X_{s,i}}\right)^2 - P_{G,i}^2} - \dfrac{V_{t,i}^2}{X_{s,i}} & \text{if } P_{G,i} \le P_{GR,i} \end{cases} \tag{20}$$

Bus Voltage Limits

$$V_i^{\min} \le V_i \le V_i^{\max}; \quad \forall i \in N_{PQ} \tag{21}$$

$$|V_i| = \text{Constant}; \quad \forall i \in N_{PV} \tag{22}$$

Security Constraints

$$S_l \le S_l^{\max}; \quad l \in N_L \tag{23}$$

$$P_{G\min,Slack} \le P_{G,Slack} \le P_{G\max,Slack} \tag{24}$$

Transformer Taps Setting Constraints:

$$T_k^{\min} \le T_k \le T_k^{\max}; \quad k \in N_T \tag{25}$$

The penalty functions corresponding to voltage violations at all load busses, reactive power violations at all generator busses, real power violations at slack bus and power flow violations at all transmission lines $\left(\lambda_{VL,i}, \lambda_{QG,j}, \lambda_{PG,Slack}\ and\ \lambda_{S,l}\right)$ are included in objective function as follows:

Generalized augmented objective function:

$$F_{aug,m} = F_{Q,m} + \sum_{i \in N_{PQ}} \lambda_{VL,i} \left(V_i - V_i^{\lim}\right)^2 + \sum_{j \in N_G} \lambda_{QG,j} \left(Q_{G,j} - Q_{G,j}^{\lim}\right)^2 + \sum_{k \in N_{G,Slack}} \lambda_{PG,Slack} \left(P_{G,k} - P_{G,k}^{\lim}\right)^2 + \sum_{l \in N_L} \lambda_{S,l} \left(S_l - S_l^{\lim}\right)^2; \quad \forall m = 1:M \tag{26}$$

Where $F_{Q,m}$ are m^{th} objective function values and the dependent variable's limiting values may be considered as:

$$V_i^{\lim} = \begin{cases} V_i^{\max}; & if \quad V_i > V_i^{\max} \\ V_i^{\min}; & if \quad V_i < V_i^{\min} \end{cases}; \qquad \forall i = 1:N_{PQ} \tag{27}$$

$$Q_{G,j}^{\lim} = \begin{cases} Q_{G\max,j}; & if \quad Q_{G,j} > Q_{G\max,j} \\ Q_{G\min,j}; & if \quad Q_{G,j} < Q_{G\min,j} \end{cases}; \qquad \forall j = 1:N_{PV} \tag{28}$$

$$S_l^{\lim} = \begin{cases} S_l^{\max}; & if \quad S_l > S_l^{\max} \\ S_l^{\min}; & if \quad S_l < S_l^{\min} \end{cases}; \qquad \forall l = 1:N_L \tag{29}$$

$$P_{G,k}^{\lim} = \begin{cases} P_{G\max,k}; & if \quad P_{G,k} > P_{G\max,k} \\ P_{G\min,k}; & if \quad P_{G,k} < P_{G\min,k} \end{cases}; \qquad \forall k = 1:N_{G,Slack} \tag{30}$$

3 HFMOEA for Solving Multi-Objective ZRPMC-VS Model

The complete details of HFMOEA algorithm develoved by Saraswat et al. are presented in reference [15]. A flowchart of HFMOEA algorithm for solving multi-objective ZRPMC-VS problem is shown in Fig.1. The details of proposed algorithm are discussed as below:

Initialization and generation of initial population : This requires input of power system data (i.e. bus data, generator data and transmission line data in as specific format) and various parameters of HFMOEA such as population size (*popsize*), maximum numbers of iterations (*max_iterations*), number of control variables, system constraints limits, initial crossover probability (P_C), initial mutation probability (P_M) etc. Initial population is generated randomly and fitness of each individual is determined.

Non-Domination Sorting: The generated initial population is sorted on the basis on non-domination sorting algorithm proposed in reference [15-16].

For producing the new population for next iteration, the following evolutionary operators are applied to parent population:

Selection: Tournament selection operator [15-16] is used for reproducing the mating pool of parent individuals for crossover and mutation operations.

Crossover: The BLX-α crossover [15] is applied on randomly selected pairs of parent individuals $\left(x_i^{(1,t)}, x_i^{(2,t)}\right)$ with a crossover probability $\left(P_C\right)$ which is a combination of an extrapolation/interpolation method.

Mutation: The PCA based Mutation [15] with mutation probability $\left(P_M\right)$ is applied to generate the offspring population.

Best compromise solution: Upon having the Pareto-optimal set of non-dominated solution using proposed HFMOEA approach, an approach proposed in [8] selects one solution to the decision maker as the best compromise solution. This approach suggests that due to imprecise nature of the decision maker's judgment, the i^{th} objective function F_i is represented by a membership function μ_i as defined in [8]:

$$\mu_i = \begin{cases} 1 & F_i \le F_i^{\min} \\ \dfrac{F_i^{\max} - F_i}{F_i^{\max} - F_i^{\min}} & F_i^{\min} < F_i < F_i^{\max} \\ 0 & F_i \ge F_i^{\min} \end{cases} \tag{31}$$

Where $F_i^{\min}$ and $F_i^{\max}$ are the minimum and maximum values of the i^{th} objective function among all non-dominated solutions, respectively. For each j^{th} non-dominated solution, the normalized membership function μ^j is calculated as:

$$\mu^j = \frac{\sum_{i=1}^{N_{obj}} \mu_i^j}{\sum_{j=1}^{N_{dom}} \sum_{i=1}^{N_{obj}} \mu_i^j} \tag{32}$$

Where N_{dom} is the number of non-dominated solutions. The best compromise solution is that μ^j which has maximum value.

Normalized fitness function: The fitness function corresponding to each individual in the population is assigned based on their respective generalized augmented functions as determined in (33). Thus the fitness function (H_n) for n^{th} objective is evaluated as:

$$H_n = \frac{K_n}{1 + F_{aug,n}}; \qquad \forall n = 1 : N_{obj} \tag{33}$$

Where N_{obj} is the total number of objectives and K_n is the appropriate constant corresponding to n^{th} objective, in this work ($K_1 = 350$ *and* $K_2 = 0.15$) .

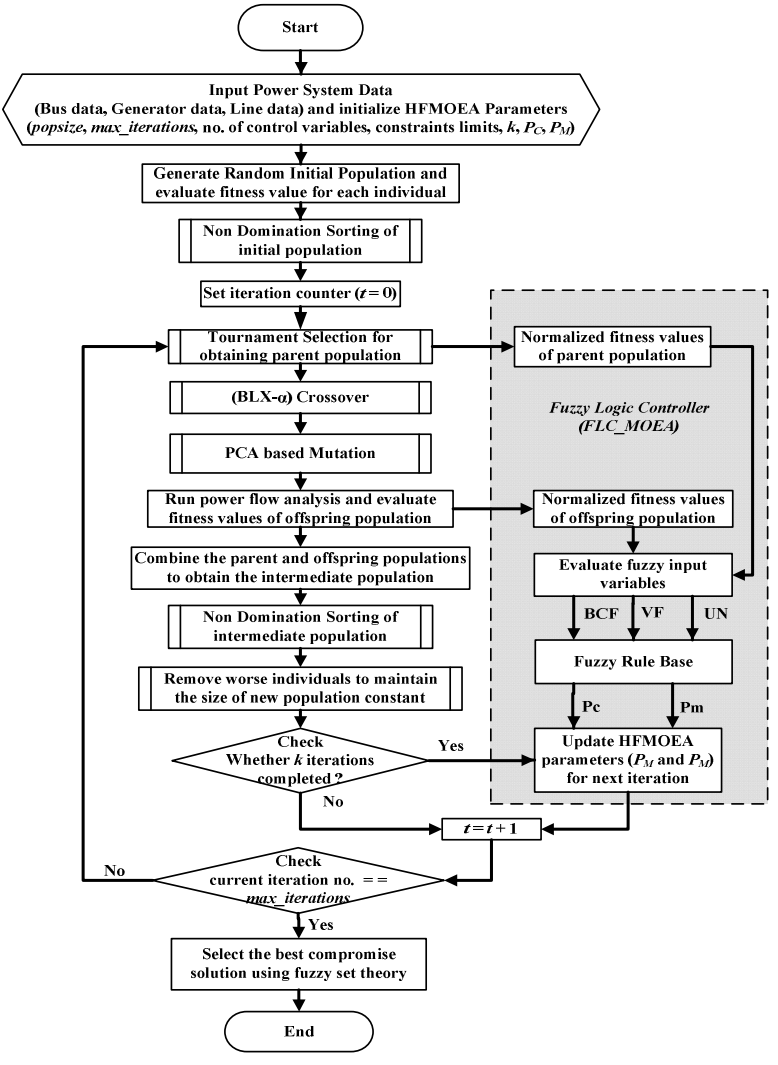


Fig. 1. Flowchart for HFMOEA for solving multi-objective ZRPMC-VS model

Fuzzy logic controller: It has been experienced that after few iterations, the fitness values of each of the individuals are becoming equal to other individuals in same population and hence the effect of crossover operator beyond that stage becomes insignificant due to lack of diversity. Therefore, the increased mutation probability (P_M) remains the only alternative to produce the better offspring for achieving a more diversified population. A fuzzy logic controller (*FLC_HMOEA*) is designed to vary P_C and P_M dynamically during the optimization process. These parameters (P_C and P_M) are varied based on the fitness function values as per following logic as in reference [15]:

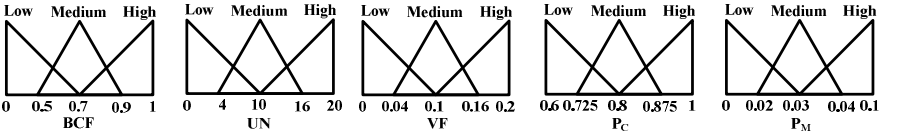


Fig. 2. Input and Output membership functions for Fuzzy Logic Controller

- The value of best compromized fitness computed (BCF) using (29)-(31) for each iteration is expected to change over a number of iterations, but if it does not change significantly over a number of iterations (UN) then this information is considered to cause changes in both P_C and P_M.
- The search for a true optimum is influenced by the diversity of a population. The variance of the fitness values of objective function (VF) of a population is a measure of its diversity. Hence, it is considered as another factor on which both P_C and P_M may be changed.

Thus the ranges of three input fuzzy parameters such BCF, UN and VF and also two output fuzzy parameters such as P_C and P_M are repersented by three lingustic terms as LOW, MEDIUM and HIGH. The details of membership functions for input and output variables of *FLC_MOEA* are shown in Fig. 2.

4 Results and Discussion

The effectiveness of HFMOEA for solving multi-objective ZRPMC-VS model is demonstrated on the IEEE 24 bus Reliability Test System (IEEE 24 RTS) (Reliability Test System Task Force, 1999) [17] shown in Fig. 3. The power system consists of 32 synchronous generators, 1 synchronous condenser (located at bus 14), and 17 constant-power type loads. The system total active and reactive loads are 2850 MW and 580 MVAr, respectively. The simulations are carried out in MATLAB 7.0 programming environment on Pentium IV 2.27 GHz, 2.0 GB RAM computer system.

In IEEE-24 RTS, system control variables are eleven generator bus voltage magnitudes, five transformer tap settings and one bus shunt inductor. Therefore, the search space has 17 dimensions. The lower and upper limits of all bus voltages are 0.95 p.u. and 1.05 p.u., respectively. The lower and upper limits of all transformer tap settings are 0.9 p.u. and 1.1 p.u, respectively. In order to carry out the RPMC-VS simulations in competitive electricity market environment, the ISO needs the following information from the reactive power providers:

Offer prices: The ISO is supposed to receive four components of the reactive power offer prices ($a_0^{i,u}$, $m_1^{i,u}$, $m_2^{i,u}$ and $m_3^{i,u}$), directly from the participants of the reactive power market. In this simulation, the reactive power offer prices (bids) submitted by all generators and synchronous condensers are taken as given in reference [18]. A synchronous condenser at bus 14 also participates in the reactive power market with its opportunity cost ($m_3^{i,u}$) equal to zero.

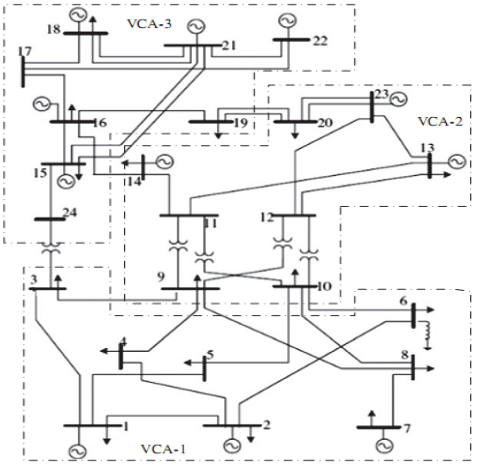


Fig. 3. IEEE 24 RTS system

Table 1. Specifications of optimization algorithms

MOEA Parameters	**NSGA-II**	**HFMOEA**
Selection	Tournament	Tournament
Crossover	SBX	BLX-α
Mutation	polynomial	PCA
P_C	0.9	varying based on FLC output
P_M	η_c=20 and η_m=20	varying based on FLC output

Generator's reactive power capability data: Each participant of the reactive power market (i.e. each generating unit and synchronous condenser) is also required to submit the information regarding its reactive power capability diagram i.e. Q_{Base} , Q_A and Q_B. In present case study, the assumptions are followed as in references [5-6] and [18] i.e $Q_{Base} = 0.10 \times Q_{\max}$, Q_A is limited either by the field or the armature heating limit, as per operating condition, and $Q_B = 1.5 \times Q_A$.

Determination of reactive power zones or VCAs: For proper management of reactive power support services, the whole power system is divided in two different reactive power zones also known as VCAs as recommended in [6]. In this paper, IEEE-24bus RTS is divided in to three zones or VCAs as shown in Fig. 3., according to the hierarchical clustering based approach as described in reference [6].

Table 2. Comparison of output results obtained after optimization for multi-objective ZRPMC model

Optimization Algorithm	**TPF**	**VSEI**	**Zone-1**				**Zone-2**				**Zone-3**			
			ρ_0	ρ_1	ρ_2	ρ_3	ρ_0	ρ_1	ρ_2	ρ_3	ρ_0	ρ_1	ρ_2	ρ_3
NSGA-II	**383.14**	**0.17164**	0.96	0	0.86	0	0.96	0	0.81	0	0.96	0	0.75	0
HFMOEA	**355.21**	**0.16931**	0.96	0	0.86	0	0.96	0	0.81	0	0.96	0	0.75	0

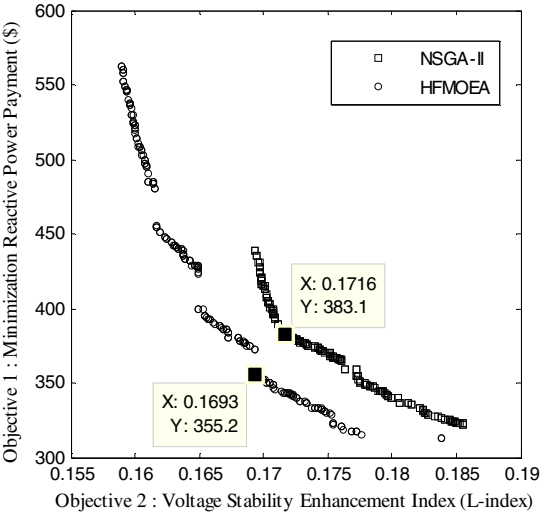


Fig. 4. Pareto-optimal fronts obtained using NSGA-II and HFMOEA in multi-objective ZRPMC-VS Models

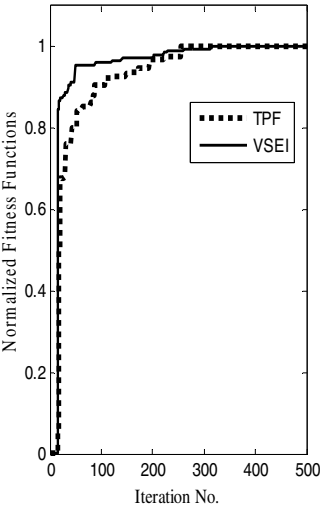


Fig. 5. convergence of TPF and VSEI in HFMOEA

The performance of HFMOEA is compared with NSGA-II [16] for solving multi-objective ZRPMC-VS model on IEEE-24 bus RTS. The detailed specifications of both the algorithms are listed in Table 1. In this case study, the similar parameters for both optimization algorithms are taken as: the number of maximum iterations (*max_iterations* = 500), population size (*popsize* = 200) and penalty factors (λ_{VLi} = 100, λ_{QGj} = 50, $\lambda_{PGSlack}$ = 50 and λ_{Sl} = 50).

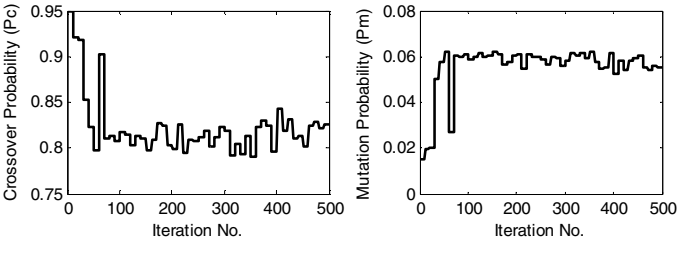


Fig. 6. Variations in crossover and mutation probabilities during HFMOEA based optimization for ZRPMC-VS problem

The optimization results obtained in solving multi-objective ZRPMC-VS model are summarized in Table 2. In multi-objective ZRPMC-VS, the best compromised solutions are selected as (**383.14$** and **0.17164**) and (**355.21$** and **0.16931**) after optimization using NSGA-II and HFMOEA respectively. The ZUMCPs $(\rho_0, \rho_1, \rho_2 \text{ and } \rho_3)$ in all three zones of reactive power market obtained

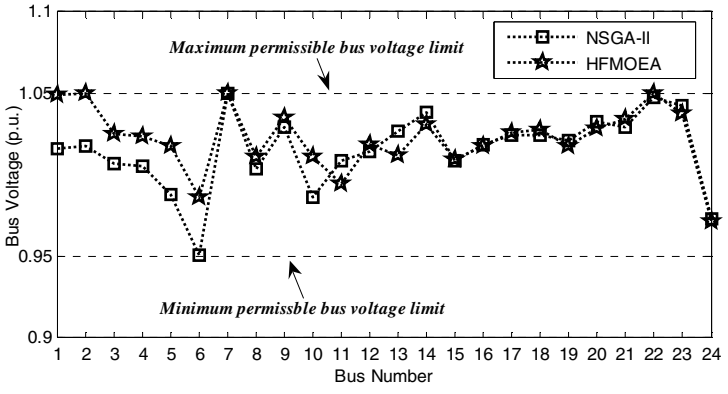


Fig. 7. Output bus voltage profiles after optimization for multi-objective ZRPMC-VS model

after execution of multi-objective ZRPMC-VS model are also mentioned in Table 2. The pareto-optimal front obtained in multi-objective ZRPMC-VS using NSGA-II and HFMOEA are compared in Fig.4. It is noticed that the pareto-optimal front and best compromised solution obtained using HFMOEA are superior compared to the same obtained using NSGA-II (see Fig.4). The convergence of both the objectives TPF and VSEI in HFMOEA based optimization process is shown in Fig. 5. It is noticed that the convergence of TPF is poor as compared to the convergence of VSEI (see Fig. 5), this is because of the higher complexity of TPF as compared to VSEI due to presence of binary variables.

The variations in HFMOEA parameters such as P_C and P_M are shown in Fig.6. It is observed that the variations in crossover and mutation probabilities are such that if P_C is going to reduce, P_M will increase (see Fig.6). As clear from Fig.4, improvement in stochastic search to reach near to true pareto-optimal front is due to P_C and P_M variations.

The bus voltage profiles obtained after optimization in multi-objective ZRPMC-VS model are compared as shown in Fig.7. It is noticed that the bus voltage profiles obtained by HFMOEA is more flat as compared to the same obtained by NSGA-II.

5 Conclusion

In this paper, a multi-objective ZRPMC-VS model is developed for improving the voltage stability of the power system while clearing the reactive power market on zonal bases. The proposed multi-objective ZRPMC-VS model is solved using HFMOEA and compared its results with the same obtained by NSGA-II. It is found that HFMOEA performance is superior to NSGA-II for obtaining better non-dominated solutions for ZRPMC-VS model.

References

1. Hao, S., Papalexopoulos, A.: Reactive power pricing and management. IEEE Transactions on Power Systems 12(1), 1206–1215 (1997)
2. Wang, Y., Xu, W.: An investigation on the reactive power support service need of power producers. IEEE Transactions on Power Systems 19(1), 586–593 (2004)
3. Hao, S.: A Reactive Power Management Proposal for Transmission Operators. IEEE Transactions on Power Systems 18(4), 1374–1381 (2003)
4. Bhattacharya, K., Zhong, J.: Reactive Power as an Ancillary Service. IEEE Transactions on Power Systems 16(2), 294–300 (2001)
5. Zhong, J., Bhattacharya, K.: Toward a Competitive Market for Reactive Power. IEEE Transaction on Power Systems 17(4), 1206–1215 (2002)
6. Zhong, J., Nobile, E., Bose, A., Bhattacharya, K.: Locallized reactive power markets using the concept of voltage control areas. IEEE Transactions on Power Systems 19(3), 1555–1561 (2004)
7. Deb, K.: Multi-objective optimization using evolutionary algorithms. John Wiley and Sons, Inc., New York (2001)
8. Abido, M.A., Bakhashwain, J.M.: Optimal VAR dispatch using a multi-objective evolutionary algorithm. Int. J. Electrical Power and Energy Systems 27(1), 13–20 (2005)
9. Zhang, W., Liu, Y.: Multi-objective reactive power and voltage control based on fuzzy optimization strategy and fuzzy adaptive particle swarm. Electrical Power and Energy Systems 30, 525–532 (2008)
10. Dai, C., Chen, W., Zhu, Y., Zhang, X.: Reactive power dispatch considering voltage stability with seekers optimization algorithm. Electric Power System Research 79, 1462–1471 (2009)
11. Jeyadevi, S., Baskar, S., Babulal, C.K., Iruthayarajan, M.W.: Solving multiobjective optimal reactive power dispatch using modified NSGA-II. Electric Power and Energy Systems 33, 219–228 (2011)
12. Rabiee, A., Shayanfar, H.A., Amjady, N.: Multiobjective clearing of reactive power market considering power system security. Applied Energy 86(9), 1555–1564 (2009)
13. Kessel, P., Glavitsch, H.: Estimating the voltage stability of power systems. IEEE Transaction on Power Systems 1, 346–354 (1986)
14. Vyjayanthi, C., Thukaram, D.: Evaluation and improvement of generators reactive power margins in interconnected power systems. IET Generation. Transmission and Distribution 5(4), 504–518 (2011)
15. Saraswat, A., Saini, A.: A Novel Hybrid Fuzzy Multi-Objective Evolutionary Algorithm: HFMOEA. In: Meghanathan, N., Chaki, N., Nagamalai, D. (eds.) CCSIT 2012, Part III. LNICST, vol. 86, pp. 168–177. Springer, Heidelberg (2012)
16. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.: A Fast Elitist Multi-objective Genetic Algorithm: NSGA-II. IEEE Transactions on Evolutionary Computation 6(2), 182–197 (2002)
17. Reliability Test System Task Force. The IEEE reliability test system – 1996. IEEE Trans. Power Syst. 14(3), 1010–1020 (1999)
18. Rabiee, A., Shayanfar, H., Amjady, N.: Coupled energy and reactive power market clearing considering power system security. Energy Conversion and Management 50, 907–915 (2009)

Comparative Study of Image Forgery and Copy-Move Techniques

M. Sridevi, C. Mala*, and Siddhant Sanyam

Department of Computer Science & Engineering
National Institute of Technology
Tiruchirappallai, India
{msridevi,mala}@nitt.edu, Siddhant3s@gmail.com

Abstract. Image forgery means manipulation of the digital image to conceal some meaningful or useful information of the image. There are cases when it is difficult to identify the edited region from the original image. The detection of a forged image is driven by the need of authenticity and to maintain integrity of the image. This paper surveys different types of image forgeries. The survey has been done on existing techniques for forged image and it also highlights various copy – move detection methods based on their robustness and computational complexity.

Keywords: Image forgery, Copy-move detection, Active approach, passive approach, Robust, Geometric transformation.

1 Introduction

Digital images play important roles in many fields such as medical, journalism, scientific publication, digital forensic etc. An image can be manipulated easily by means of an image editing tool such as Photoshop. Manipulations of digital images are done for hiding some meaningful or useful information, to create misleading images or to make forged images [2]. Figure 1 gives an example, in which it is difficult to identify forensic / manipulated image Figure 2 from the original one. The image forensics aims to address image authenticity and integrity. Image tampering, splicing or cloning has been done to create forged images. Therefore the integrity of the image is lost. The digitally forged images are sometimes so real and it cannot be distinguishable from the original image. Hence authenticity is also lost. Integrity and authenticity verification of digital images are one of the hot and serious research issue in the field of image processing. During past few years many research papers have been published in this area. Already proposed methods are reviewed in this paper with respect to their computational complexity and reliable detection method (i.e) robust to some common process such as compression, scaling, rotation, translation, noise etc. Even though there are many methods available for detecting digital image forgery, the success of the method is limited to some extent.

* Corresponding author.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 715–723.
springerlink.com

Fig. 1. Original Image

Fig. 2. Forged Image

The most frequently used technique for image tampering is copy-move which aims to hide or manipulate the content of the image. This paper surveys the different approaches applied for detecting copy-move forgery.

The paper is organized as follows: Section 2 discusses about the classification of image forgery approaches and drawbacks of them. Section 3 explains different techniques proposed for image tampering. Section 4 describes the comparison of different methods based on their complexity and robustness. Final Section 5 concludes the paper.

2 Image Forgery Approaches

The image forgery approaches are basically classified into two types namely active and passive approaches [1]. Both approaches use different techniques and they are classified further [23] which is shown in Figure 3.

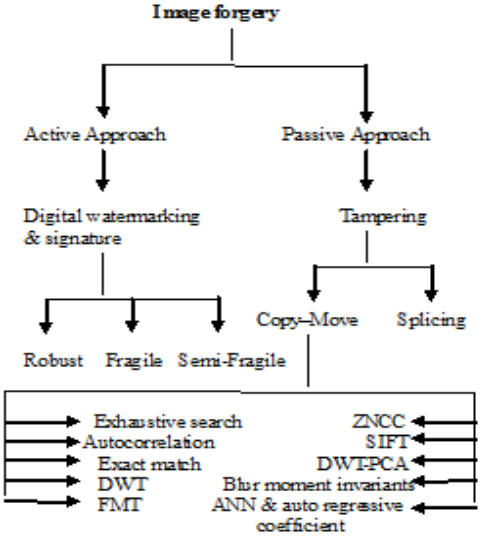


Fig. 3. Classification of image forgery approaches

(a) Active approach: The active approaches are mainly based on digital watermarking and signature as explained in [3, 23]. Watermarking can be performed either in spatial or frequency domain. In spatial domain, watermark is directly embedded into the pixel such as LSB (Least Significant Bit). The problem with this method is easy detection of watermarked data. Frequency domain enhances better data security when compared to spatial domain due to its complex calculations. So, this technique involves conversion of image from spatial to frequency domain by use of transform such as Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), Discrete Fourier Transform (DFT) etc. Then the watermark symbol is embedded into the frequency domain coefficients.

Water marking techniques are classified further based on robustness, fragile and semi fragile referred in [3].

i) Robust watermarking: Watermark stands with the image even after processing such as translation, rotation, scaling and compression are applied to the image. This method is used to protect the copy right of digital media.

ii) Fragile watermarking: The slight change or modification on the image may lead to invalid image. By this, authenticity of image can be verified.

iii) Semi fragile watermarking: It is intermediate between first and second method. It can distinguish between malicious (modification, cropping etc.) and non malicious attacks (compression, smoothening etc.)

***(b) Passive approach*:** The passive approach does not rely on pre-registration/pre-embedded information but uses the image processing techniques for authenticity detection. Passive blend/digital forensics is basically in any one of the following forms.

- Tampering
- Splicing
- Cloning

Table 1. Review of image forgery approaches

S.No	Approach	Techniques	Characteristics	Merits	Demerits
1.	Active approach	Digital Watermarking & Signature	i) Watermark symbol has to be embedded into the original image ii) Use Cryptographic technique	Low Computational complexity compared to passive approach	- Need Preprocessing of image - Does not locate modified part
2.	Passive approach	Copy – Move	Copied Segment is pasted anywhere in the same image.	- Do not rely on pre-embedded information - Uses image processing technique. - Difficult to identify the pasted part	Detection involves high computational complexity
		Splicing	Two or more region of the images combined into a single digital image		

3 Image Forgery Methods

3.1 Digital Watermarking

The procedure behind watermarking is shown in Figure 4. The method proposed in [3] provides authentication for JPEG images based on Genetic Algorithms. The original image is divided into 8 x 8 blocks and mark each block b(i,j), i^{th} number X. The new mapping block number X^1 is formed by applying transform to the block number X. The authenticate information are generated using CRC (Cyclic Redundancy Check) for each block b(i,j) and its values are adjusted using Genetic algorithms (Substitution of chromosome in the original image, adjust the block), so the modified block b^1 (i,j) after JPEG compression contains the authenticated information. All modified blocks b^1 (i,j) are combined to generate a modified image B. The watermarked image C is obtained by compressing the modified image B using JPEG compression with QF (Quantization Factor) [16].

3.2 Copy -Move Detection

The steps involved in copy – move detection is shown in Figure 5. The first step in copy – move detection is dividing the image into overlapping blocks and then constructs a matrix for the overlapping blocks [15, 17]. The constructed matrix is sorted (Lexicographical [18] / Kd tree [19]). The duplicated regions are adjacent to each other in the sorted matrix. The existing methods differ both in their computational complexity and robustness.

Exhaustive search method explained in [6], is simplest and most obvious approach. It finds out the closely matching segments of the image and its circularly shifted version. The copied region can be calculated by the formula given in the paper [6] by comparing X_{ij} (pixel value of a image at the position (i,j)) with its cyclical shift [k,L], which is same as comparing the value with its cyclical shift $[k^1,L^1]$ where $k^1=M-k$, $L^1= N-L$. Even though this method is simple and effective, it is not used quietly due to its computational complexity. The computational requirement is $(MN)^2$ for the image of size M x N. The next method uses auto correlation. The copied and pasted segments will produce peaks for autocorrelation of the shifts. The autocorrelation of the image X of the size M x N is given in [6] as

$$R_{k,L} = \sum_{i=1}^{M} \sum_{j=1}^{N} X_{i,j} X_{i+k,j+L} \quad \begin{array}{l}\text{where } i,k = 0,1\ldots.M-1 \\ \text{and } j,L = 0,1,\ldots N-1\end{array}$$

Another approach is exact match or block match [4]**.** This approach is significantly better and faster than the above two methods. The block size selection is the major problem in block matching method because detection depends upon the size of block. If block size is large then it will lead to failure even though the copied part is present on it. If block size is too small, it appears false positives.

The methods discussed above will work only for the forged image which does not undergo any geometric transformations. It will detect properly if the copied part is pasted over the original image as such without any modification to the copied part.

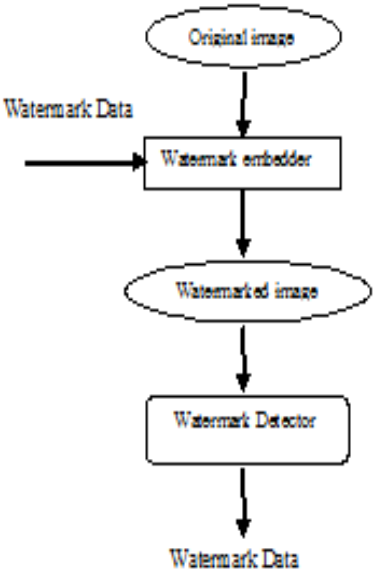


Fig. 4. General Procedure for Watermarking

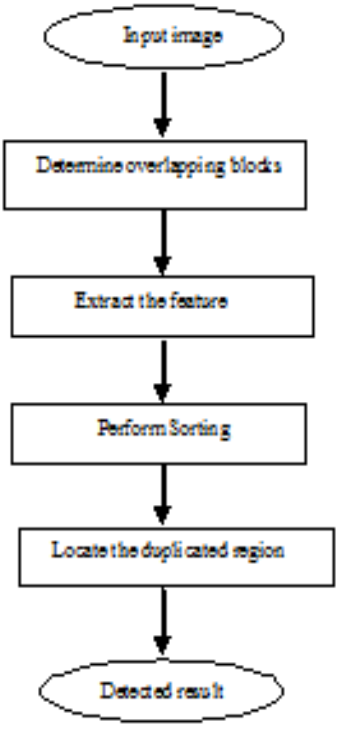


Fig. 5. Basic Steps of Image Tampering

To overcome the problems of the above methods, robust detection method [12] is used. The robust method for identifying forged image is by using Discrete Wavelet Transform (D WT) as explained in [11]. The detection is performed in reference and matching blocks on the lowest level of wavelet transform compressed image. The more robust detection can be attended by checking with different DWT levels. Phase correlation is used for checking similarity region present in the original image. Next method deals robustness by making use of Fourier Mellin Transform (FMT) [7] for finding tampered image. It properly detects the copied region only if the region undergoes geometric transformations. The image is divided into B x B overlapping blocks. Consider a block i(x,y) and its geometric transformation version i^1 (x,y) where i^1 (x,y) = i (σ (xcosθ + ysinθ) – x_0, σ (y cosθ - xsinθ) - y_0). Here (x_0, y_0) is translation factor, σ is scaling factor and θ is angle of rotation.

The method discussed in [20] for finding copied region is by applying Zero Mean Normalized Cross Correlation (ZNCC). Consider the given input image of size M x N and it is divided into number of overlapping blocks [NB] = (M-B+1) * (N-B+1) with each B x B size. In order to save the memory space of the pixel intensity data, store only top left pixel of each block into an array. The pixel intensities of the block are retrieved from the image whenever they are needed. The computational cost is reduced using (k-Dimensional) kd– tree sorting [19]. The computational complexity of kd – sorting is low, [O(NB X log2 (NB/Nss))] where NB is number of blocks, Nss is Neighborhood

Search Size. The matching of the duplicated region takes complexity of [O(NB X Nss)]. The algorithm complexity is O(NB X Nss) where Nss < N. Normalized cross correlation function is used for similarity measurement. The method is robust to match in presence of minor noise and lossy compression. The cross correlation between each block B_i and B_j is computed, where j=i+1,… i+Nss. If the result of the computed value is less than the specified threshold or previously found maximum ZNCC value, the pair is ignored. Thus it allows the best matches within the Nss neighborhood and discards the rest. If the duplicated region is present in the image, then the copied source and pasted destination will appear as two monochromatic clusters of pixels. The method does not work if rotation and scaling are applied to the copied part.

A new method was proposed in [21], [22] for detecting identical region in presence of post copy paste operations like blurring and adding noise based on Blur moment invariants. Consider 2D (p + q)th order central moment for image f(x,y) is

$$\mu_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-x1)\,(y-y1) f(x,y)\, dxdy$$

The proposed method uses 24 blur invariants for gray scale image and 72 for color image upto 7^{th} order to create feature vector of each block. It uses principle component transformation to reduce the feature vector dimensions. The algorithm uses kd-sorting for determining similar blocks. The duplicated region is calculated based on Eclidean distance among blocks.

In [9], the authors suggested a method using Discrete Wavelet Transform – Principle Component Analysis (DWT-PCA) which accurately detect cloned image as long as the copied segment is not scaled or rotated. The algorithm divides the image into 4 sub bands using DWT. The PCA [5] – Eigen value decomposition (EVD) on each of the k1 row vectors is performed to reduce vector length to $t < b^2$. (Form a matrix A containing k1 x t elements). The rows of the matrix are sorted lexicographically. Duplicated regions are adjacent to each other in the sorted matrix.

The approach based on Artificial Neural Network (ANN) and Auto Regressive Coefficient is explained in [10]. The forged image has high correlation in digital image data. The identical part in the image can be formulated using auto regression coefficient (as feature vector). The feature vectors [13] of different images are trained on ANN. For training the Neural Network, 2 classes of feature vectors are collected (one group from the original images and other from the forged images). The author concluded that the forged images can be identified properly by training the forged images in ANN in a better way. So it will perform best detection for rotation, scaling and resizing applied to the pasted region.

The next method based on SIFT (Scale Invariant Feature Transform) [1]. The method extracts distinctive features of image which are invariant to scaling, rotation, and robust to noise, illumination, distortion. The forgery of the image is detected by feature matching process. The matching for each keypoint is identified from its neighbor.

3.3 Image Splicing

An algorithm to verify authenticity of image quality features like Markov and moment based features is explained in [8]. These approaches produce best result for the

image splicing [14]. The basic step in this approach is extraction of image quality metrics (IQMs) & moment based features. The next step is to calculate all the subbands and then obtain histogram for them, determine its characteristics function by applying DFT to the histogram. Finally moments are calculated. Based on the calculations, find out the best parameter for training the model and obtain SVM model.

4 Comparsion of Computational Complexity of Different Copy – Move Algorithms

The various copy-paste region detection techniques are explained in section 3. The comparisons of those methods are tabulated in Table 2 with its computational com

Table 2. Comparison of copy– move detection methods

S. No	Method	Characteristics	Computation involved	Computational Complexity
1.	Exhaustive search	Does not work for post processing of copy – paste region	Searching	$(MN)^2$
2.	Autocorrelation	Does not work for post processing of copy – paste region	Autocorrelation	$< (MN)^2$
3.	Exact match	Does not work for post processing of copy – paste region	Lexicographical sorting	$MN \log_2(MN)$
4.	Robust match (DWT)	Work for noise addition and change in JPEG quality level, but does not work well under rotation and scaling.	Search using Pyramid	Lower complexity (Search performed only on the lower resolution)
5.	Robust match (FMT)	Work for Geometric transformation such as Translation, scaling and only for some small angle of rotation.	Counting bloom filter	MN
6.	ZNCC	Does not work for rotation and scaling.	Kd – sorting	NB X Nss
7.	Blur moment invariants	Work for Compression like JPEG	Blur invariants based on central moments	Higher complexity
8.	DWT-PCA	Accurately detect if the copied segment is not scaled or rotated	Lexicographical sorting	O(8k log k) for DWT
9.	Artificial Neural Network and Auto Regressive Coefficient	Detect well for postprocessing such as Rotation, scaling, resizing of copy-move part	Auto regressive coefficient	Higher complexity (Training)
10.	SIFT	Find properly for Scaling, Rotation, Additive noise and compression part of the image	Kd–tree Sorting	Higher complexity (due to 128 elements)

plexity along with its characteristics. Consider M x N is the size of the input image, NB denotes number of overlapping blocks (each of size B x B), Nss is local neighborhood search size, $k = (N-b+1)^2$.

5 Conclusions

Many techniques are available for identifying digital image forgery such as tampering, splicing, cloning etc. Some of them were discussed in this paper. The main problem in copy – move digital image forgery detection is selection of block size. If the block size is too small, then it leads to false positive appearance. If the block size is too large then some forged areas remain undetected. Hence there is a need to select an optimal block size for proper detection of forged image. The computational complexity of the algorithms mainly depends on selection of block, sorting and searching techniques applied on it. In future, the time, space complexity and robust of the image forgery algorithm can be improved by means of reducing the size of the image by using compression techniques to make search faster and robust as possible for post processing.

References

[1] Huang, H., Guo, W., Zhang, Y.: Detection of Copy-move forgery in digital image using SIFT algorithm. In: IEEE Pacific–Asia Workshop on Computational Intelligence and Industrial Application, pp. 272–276. IEEE computer society (2008)

[2] Kang, L., Cheng, X.-P.: Copy – move forgery detection in digital image. In: 3rd International Congress on Image and Signal Processing (CISP), pp. 2419–2421 (2010)

[3] Edupuanti, V.G., Shih, F.Y.: Authentication of JPEG Images based on Genetic Algorithms. The Open Artifical Intelligence Journal 4, 30–36 (2010)

[4] Ardizzone, E., Bruno, A., Mazzola, G.: Copy – move forgery detection via texture description, pp. 59–64. ACM (2010)

[5] Yang, Q.-C., Huang, C.-L.: Copy – Move Forgery detection in digital image. Springer, Heidelberg (2009)

[6] Fridrich, J., Soukal, D., Lukas, J.: Detection of copy- move forgery in digital images. In: Proceedings of Digital Forensic Research Workshop, Cleveland (August 2003)

[7] Bayram, S., Sencar, H.T., Memon, N.: An efficient and robust method for detecting copy-Move forgery. In: International Conference on Acoustics, Speech and Signal Processing, pp. 1053–1056 (2009)

[8] Math, S., Tripathi, R.C.: Image quality feature based detection algorithm for forgery in images. International Journal of Computer graphics and animation (IJCGA) 1(1), 13–21 (2011)

[9] Zimba, M., Sun, X.: DWT- PCA (EVD) based copy – move image forgery detection. International Journal of Digital Content Technology and Its Applications 5(1), 251–257 (2011)

[10] Gopi, E.S., Lakshmanan, N., Gokul, T., Kumaraganesh, S., Shah, P.R.: Digital image forgery detection using artificial neural network and auto regressive coefficients. In: IEEE CCECE/CCGEI, pp. 194–197 (2006)

[11] Khan, S., Kulkarni, A.: Robust Method for detection of copy- move forgery in digital images, pp. 69–73. IEEE (2010)

[12] Kumar, S., Da, P.K., Shally: Copy – Move Forgery detection in Digital Images: Progress and challenges. In: International Conference on Computer Science and Engineering (IJCSE), vol. 3(2) (February 2011)
[13] Christlein, V., Riess, C., Angelopoulou, E.: A study on features for the detection of copy – move forgeries
[14] Farid, H.: Image Forgery Detection – A Survey. IEEE Signal Processing Magazine, 16–25 (2009)
[15] Bayram, S., Sencar, H.T., Memon, N.: A survey of copy-Move forgery detection techniques. In: IEEE Western New York Image Processing Workshop, New York (September 2008)
[16] Shih, F.Y., Yuan, Y.: A comparison study on copy – cover image forgery detection. The Open Artificial Intelligence Journal 4, 49–54 (2010)
[17] Shivakumar, B.L., Santhose Baboo, S.: Detecting copy – move forgery in digital images: A survey and analysis of current methods. Global Journal of Computer Science and Technology 10(7), 61–65 (2010)
[18] Wiedermann, J.: The complexity of Lexicographic sorting and search. Computing Research Center
[19] Talbert, D.A., Fisher, D.: An empirical analysis of techniques for constructing and searching k-dimensional trees. In: International Conference on Knowledge Discovery and Data Mining, pp. 26–33 (2000)
[20] Jung, I.K., Lacroix, S.: A robust interest point matching algorithm. In: International Conference on Computer Vision (2001)
[21] Mahdian, B., Saic, S.: Detection of copy – move forgery using a method based on blur moment invariants. In: Forensic Science International Conference, pp. 180–189 (2007)
[22] Flusser, J., Suk, T., Saic, S.: Image features invariant with respect to blur. Patten Recognisation, 1723–1732 (1995)
[23] Cox, I.J., Miller, M.L., Bloom, J.A.: Digital watermarking principles and practices (2002)

Single Sideband Encoder with Nonlinear Filter Bank Using Denoising for Cochlear Implant Speech Processor

Rohini S. Hallikar[1], M. Uttara Kumari[1], and K. Padmaraju[2]

[1] Department of ECE , R.V. College of Engineering , Bangalore, India
rohini_hallikar@yahoo.co.in
[2] Departments of ECE JNT University, Kakinada, Andhra Pradesh, India

Abstract. Cochlear Implants (CI) are the most successful neural prosthesis used to restore normal hearing to the profoundly deaf, by electrical stimulation of the auditory nerves. The use of speech coder is very crucial in the cochlear implant to obtain a very close resemblance of the normal hearing. Use of noise reduction techniques further enhances a satisfactory hearing in noisy conditions. We propose a new method of sound processing which gives improved speech recognition. To achieve this goal we implemented denoising technique and further adopted SSB demodulation along with a non linear filterbank such as The Dual Resonance Non Linear (DRNL) which is capable of modeling the behavior of the human cochlea. Comparative analysis was done to understand the performance of the proposed method with existing method. Simulation results showed a significant improvement in the speech recognition over existing method.

Keywords: Neural Prosthesis, Cochlear Implants, DRNL, Denoising.

1 Introduction

Individuals with profound impairment can look towards cochlear implants as a means of obtaining hearing capabilities. Basilar membrane uses a tonotopic organization. This basically refers to a mapping process, which is referred to as frequency-place mapping. The hair cells located along the basilar membrane responds differently to different frequencies. The brain is later capable of locating these responses and accordingly we hear the sound. In individuals with profound deafness, the hair cells are very less .Cochlear implants bypasses these hair cells and can directly stimulate the cochlear nerves using electric stimulations.

One of the important goals of speech processing methods in cochlear implants is to obtain a good mimicking of the natural hearing capabilities. Choice of a speech processing method is very crucial to get better speech hearing. Researchers are constantly working to obtain better speech processing methods to obtain the above goal. Many cochlear implants uses encoded coarse features discarding the fine structure.

Intially fundamental frequency and second resonance frequency were used in case of multi-electrode Nucleus 22.

Later versions, the first formant were added, followed by three spectral peaks between 2000 and 8000 Hz. Consistent improvement was observed as more spectral

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 725–734.
springerlink.com

details were added. Later methods used temporal envelopes instead of spectral envelopes and high level of speech recognition was observed.

Encoding of spectral and temporal fine structure cues in cochlear implants is the focus of present signal processing methods.

Increasing the electric stimulation carrier rate and extraction of frequency modulation from the temporal fine structure are some of the methods of encoding fine structure And in another technique we may use multiple carriers to encode the fine frequency structure [1]. An improved CI speech processing strategy using a time varying filter model of a biological cochlea offers many advantages. Notable improvement in performance is a result of robust formant representation in noisy conditions [2].

This paper discusses an improved method of speech processing which incorporates non-linear techniques for speech identification.SSB demodulation makes it feasible to obtain a better performance in melody recognition A combination of speech processing method making use of non-linear filter bank and SSB demodulation technique along with denoising would result in better speech recognition. One of the parameter for measurement of performance of the proposed method is by using correlation coefficient. Alternatively a weighted mean correlation coefficient can be computed using a Fisher transformation [3]. DRNL is a good technique which offers the advantages of being an accurate cochlear model and having a computational simplicity [4].

2 Existing Techniques

Some of the prevalent speech strategies are compressed analog, continuous interleaved sampling, Spectral PEAK estimation & Amplitude time frequency, nonlinear filter banks.

2.1 Compressed Analog (CA)

This was the basic vocoder –centric strategy, which was one of the first techniques to be used in cochlear implants. It consisted of an analyzer and a synthesizer. The signal, to be transmitted, was compressed using automatic gain control, and then filtered into four contiguous frequency bands. These filtered waveforms were transmitted to four intracochlear analog form electrodes. The simultaneous stimulation, however, caused interactions between the channels and hence distorting the spectral information [5].

2.2 Continuous Interleaved Sampling (CIS)

Continuous Interleaved Sampling makes use of the speech strategy methodology in its simplest form, making use of pulsatile stimulation as described in [6]. It makes use of an 8 channel filter-bank, which is fixed. It extracts the envelope, and uses this envelope to modulate a biphasic pulse train.

2.3 SPEAK

SPEAK stands for Spectral PEAK extraction. It follows a similar methodology compared to CIS, with the additional implementation of N of M logic which chooses the

N highest filter bank contributors out of M. This strategy employs a trade off, a drop in speech quality for an increase in the stimulation rate.[7] SPEAK is very similar to CIS, except in the choice of choosing the N channels out of M which contribute most to the reconstructed sound signal and the increased stimulation rate which results in better intelligibility but reduced spectral quality.

2.4 ATF(Amplitude-Time-Frequency)

It is a coding technique employed for the pulsatile series generated after the filter bank to achieve higher stimulation rates & lower the number of electrodes to be implanted, thus lowering the cost [8].

2.5 Nonlinear Filter Bank Model

This processing strategy is based on a nonlinear time-varying filter model of a biological cochlea. The level dependent frequency response characteristic of the basilar membrane is known to produce robust formant representation and speech perception in noise. A Dual Resonance Non Linear (DRNL) model is simpler than other adaptive non-linear models of the basilar membrane [2].

In general, the input signal is first pre emphasized. The entire frequency range of the preemphasized signal is then decomposed to eight bands of frequency by passing it through a DRNL based frequency decomposition stage consisting of eight channels. The envelopes of the outputs of the channels are later compressed and used to modulate biphasic pulses. Compression functions making use of non linearity of logarithmic type have the advantage of the envelope fitting the patient's dynamic range of hearing. The envelopes are further processed to obtain biphasic pulses. The amplitude of these pulses is dependent upon the envelopes. These biphasic pulses are fed to the electrodes so as to avoid overlapping. Also the rate at which they are fed to the electrodes is constant.

2.6 Single Sideband Encoder

Here an audio signal is converted into a time - varying electrically stimulating pulse trains. The sound is first split into several frequency subbands and each sub band signal is coherently downward shifted to a low-frequency base band. These resulting coherent envelope signals are real-valued. Using a peak detector each sub band is further converted into rate-varying and interleaved pulse trains [9].

3 Proposed Method

3.1 Single Sideband Encoder with Non Linear Filter Bank Using Denoising

This method uses a combination of SSB Encoding and a DRNL filter bank. The goal here is to obtain good speech recognition in noisy conditions. The signal from the microphone is first denoised, followed by the preemphasis stage. The next stage is the DRNL filter block. Here, instead of dividing the signals into fixed linear band pass

channels, the input speech is divided into multiple sub-bands by the DRNL filter array. The output of each of the sub-band is then passed through an SSB encoder for extraction of temporal cues. The envelope of the signals are extracted and compressed logarithmically by mu-law compression. The envelope is used to modulate a biphasic pulse train. Finally the reconstructed signal is compared with the original signal by taking the correlation into account. Next we compute the correlation coefficient using eq(1). Figure 1 shows the implementation of the proposed scheme.

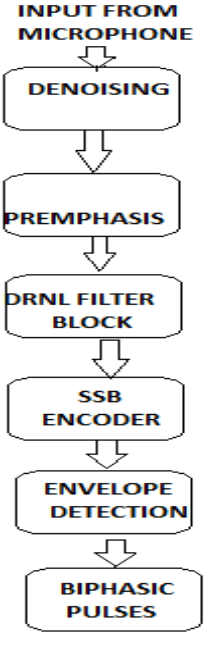


Fig. 1. Block diagram of proposed method

4 Results

For various speech input signals, the correlation coefficient was computed with the proposed method, SSB with DRNL & DRNL methods. The correlation coefficient between the input signal (x) and reconstructed signal (y) was computed by using eq(1).

$$r = \frac{\sum_i (x_i - x')(y_i - y')}{[\sum_i (x_i - x')^2 (y_i - y')^2]^{0.5}} \quad (1)$$

Where, r is the correlation coefficient between two signals x and y, with x' and y' representing their means.

Table 1, shows the simulation results for news.wav for different listening conditions.

Three different algorithms were simulated to check their performances. The first method was the proposed method, which used the SSB with DRNL method along with the denoising stage. The second was the SSB with DRNL and the last method was DRNL. From Table 1, it is observed that SSB with DRNL, proposed method gives better result compared to DRNL method especially under noisy condition. Under quiet condition, there is an improvement of 2.28% and 2.64 % with proposed method and SSB with DRNL respectively over DRNL. And also there is a significant improvement in the correlation coefficient under noisy conditions with the proposed method. The improvement for proposed method with 10dB noise is 15.46 % over DRNL.

Table 2, 3 and 4 shows the simulation results for different speech signals, for different sampling frequencies taking into account, different listening conditions for three algorithms.From Table 2, it is observed that with the increase in the sampling rate to 20 KHz, could still give good results of proposed method with respect to SSB with DRNL and DRNL methods. But compared to the same speech wav file sampled at a lower sampling rate the increase in performance for quiet conditions was 1.12%. For 10 dB babble noise, the increase in performance was 18.7%. This clearly shows the superiority of the proposed method under noisy condition. Simulation results shown in fig 2 to fig 4 indicate that the proposed method is superior to the SSB with DRNL, and the DRNL method for cases of noiseless as well as noisy inputs.

Increasing sampling frequency from 10 kHz to 20 KHz does not help to get a better speech output in both noisy and noiseless conditions. Hence 10KHZ is the optimum value required to get good results. The proposed method showed an improvement of 8.12% compared to DRNL for a higher sampling frequency, i.e. Fs=20 KHz, for 10dB babble noise as given in Table 4.

The waveforms of clean speech files and its formant frequencies are represented in Fig 5 & 6 respectively. Table 5 shows that out of the five formant frequencies representations for flower.wav (clean signal), three formant frequencies are having higher values compared to two formant frequencies for news.wav (clean signal). And also it is observed that the correlation coefficient was higher for flower.wav compared with news.wav signal for noisy as well as noiseless speech signals.

Table 1. Comparison of correlation coefficient for three different methods Fs=10KHz for the speech signal 'news.wav'

Table .	Proposed Method (1)	SSB with DRNL (2)	DRNL (3)
Quiet	0.8096	0.8068	0.7888
5 dB babble noise	0.4870	0.4857	0.4377
10 dB babble noise	0.5505	0.5488	0.4768

Table 2. Comparison of correlation coefficient for three different methods Fs=20 KHz for the speech signal 'news.wav'

Table .	Proposed Method	SSB with DRNL	DRNL
Quiet	0.8006	0.7978	0.7658
5 dB babble noise	0.4757	0.4745	0.4136
10 dB babble noise	0.5404	0.5387	0.4562

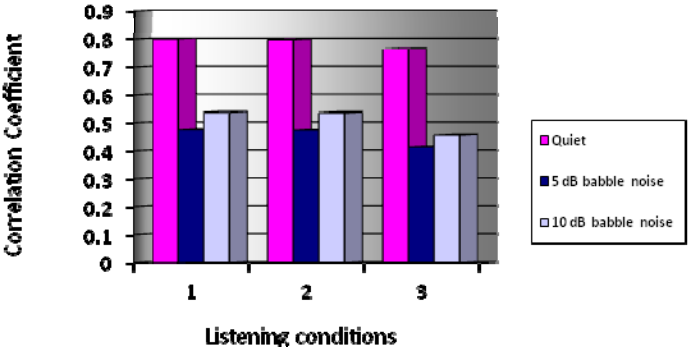


Fig. 2. The performance of proposed method, SSB with DRNL & DRNL methods for different listening conditions for news.wav using Fs=20KHz

Table 3. Comparison of correlation coefficient for three different methods Fs=10KHz for the speech signal '**flower.wav**'

Listening Conditions	Proposed Method	SSB with DRNL	DRNL
Quiet	0.8379	0.8352	0.8379
5 dB babble noise	0.6211	0.6198	0.5618
10 dB babble noise	0.7033	0.7017	0.6560

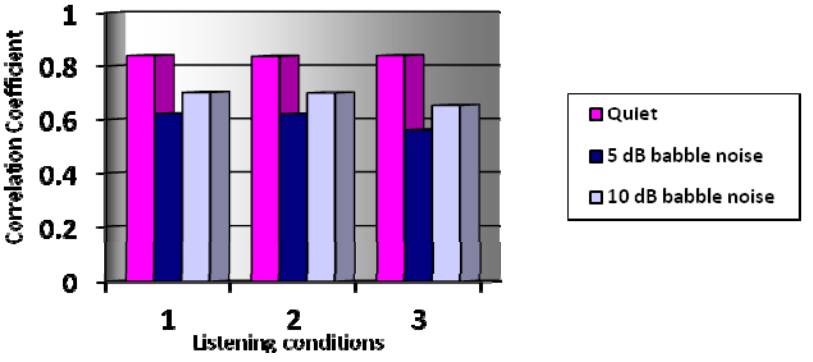


Fig. 3. The performance of proposed method, SSB with DRNL & DRNL for different listening conditions for news.wav using Fs=**10KHz**

Table 4. Comparison of correlation coefficient for three different methods Fs=20KHz for the speech signal '**flower.wav**'

Table .	Proposed Method	SSB with DRNL	DRNL
Quiet	0.8231	0.8204	0.8111
5 dB babble noise	0.6154	0.6141	0.5510
10 dB babble noise	0.6953	0.6937	0.6431

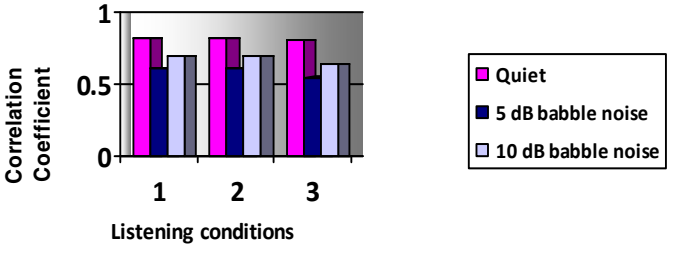


Fig. 4. The performance of proposed method with SSB with DRNL and DRNL methods for different listening conditions for news.wav using Fs=20KHz

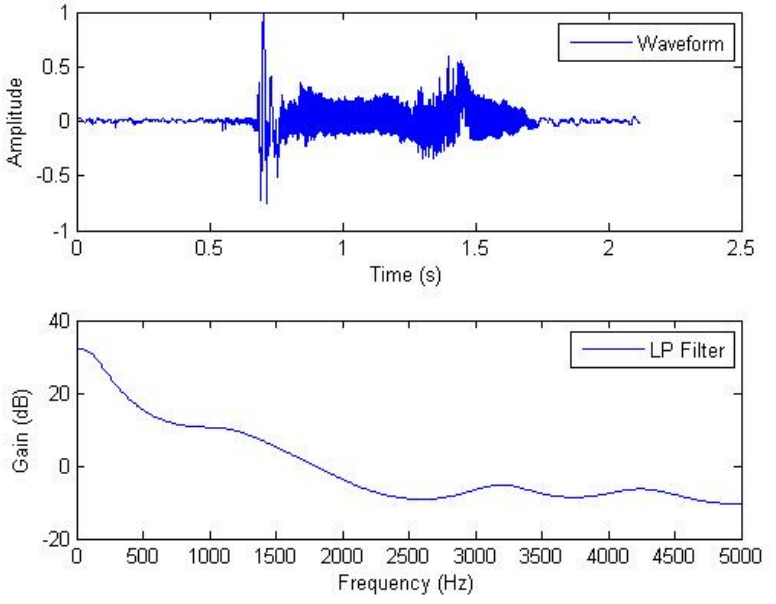


Fig. 5. Clean speech signal & formant frequencies for flower.wav

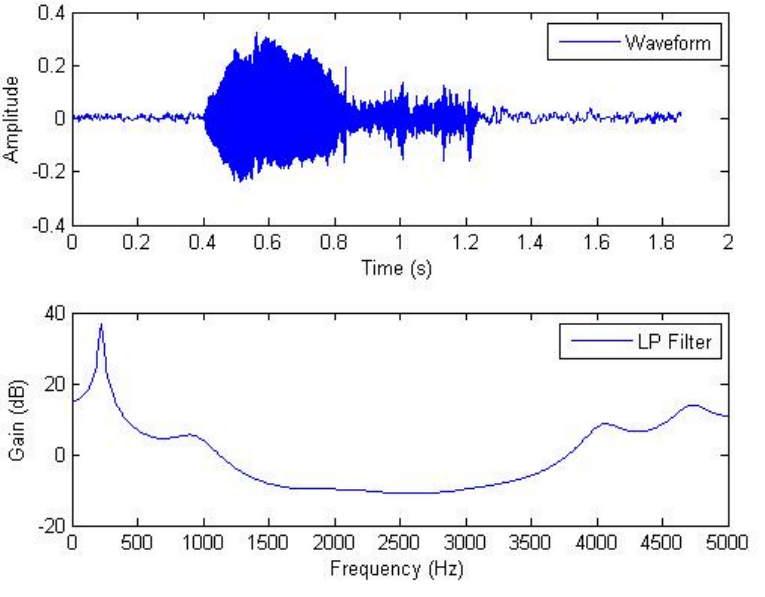


Fig. 6. Clean speech signal & formant frequencies for news.wav

Table 5. Comparison of formant frequencies and correlation coefficient for different speech signals, Fs=10 KHz under quiet condition

Table .	Formant 1	Formant 2	Formant 3	Formant 4	Formant 5	Formant 6	Correlation Coefficient
flower	104.1	1168.3	1680.4	3191.5	4253.3	-	0.8379
News	223.5	941.5	2094.8	3105.7	4038.7	4718.0	0.8096

5 Conclusion

The key contribution of this paper is the implementation of a new method which results in a better speech. Correlation coefficient is an indicator of the signal fidelity, and in this case considering the realistic condition of noisy signals, the performance of proposed method is superior to the existing method.SSB with DRNL, proposed method shows an improvement in the reconstructed waveform under noisy conditions over the existing method. The two algorithms were found to perform better under noisy conditions.

References

[1] Zeng, F.-G. IEEE Senior member, Rebscher, S., Harrison, W., Sun, X., Feng, H.: Cochlear implants: system design, integration, and evaluation. A Clinical Application Review, IEEE Reviews in Biomedical Engineering 1, 115–142 (2008)

[2] Kyung, H.K., Sung, J.C., Jin, H.K., Doo, H.K.: An improved speech processing strategy for cochlear implants based on an active nonlinear filterbank model of the biological cochlea. IEEE (2009)

[3] Garcia, E.: A Tutorial on Correlation Coefficients (July 8, 2010) (last modified on January 8, 2011)

[4] Marracci, M., Tellini, B., Lazzarino, L.L., Giannetti, R.: On the use of the Dual Resonance Non Linear filter for speech processing in a Hearing Aid Algorithm

[5] Loizou, P.C.: Speech processing in vocoder-centric cochlear implants, pp. 109–143 (2006)

[6] Wilson, B.S., Niparko, J.K. (eds.): Strategies for representing speech information with cochlear implants, in cochlear implants: principles and practices, pp. 129–170. Lippincott Williams & Wilkins, Philadelphia (2007)

[7] Nogueira, W., Buchner, A., Lenarz, T., Edler, B.: A psychoacoustic "NofM"-type speech coding strategy for cochlear implants published in the EURASIP. Journal on Applied Signal Processing (2005)

[8] Chen, W.-B., Zhou, L.-H., Xiao, Z.-J., Chen, G.-J., Wang, L.-J.: A new speech coding for improving the quality of cochlear implant. BioMedical Engineering and Informatics (2008)

[9] Nie, K., Atlas, L., Rubinstein, J.: Single sideband encoder for music coding in cochlear implants. Departments of Otolaryngology, Bioengineering, and Electrical Engineering Virginia Marrill Bloedel Hearing Research Center, University of Washington, Seattle, USA. IEEE (2008)

Authors

Rohini S. Hallikar Completed B.E. (Electronics) degree from Dr.B.A.M.U Aurangabad, Maharashtra. Completed M.Tech (Digital Electronics & Communication),VTU, Belgaum. Currently working as Assistant Professor in department of Electronics and Communication, R.V. College of Engineering, Bangalore. 560059, Karnataka, India

M. Uttara Kumari received the B.E degree in 1989 from Nagarujna University, Hyderabad, Andhra Pradesh and M.E degree in 1996 fromBangalore University, Karnataka and Ph.D degree in 2007 from Andhra University. Presently working at R.V. college of Engineering with an experience of 17 years in the teaching field. Her research interest lies in various areas of radar systems, Space-time adaptive processing, speech processing and image processing.

Dr. K. Padma Raju Completed B.Tech (E & C) from Nagarjuna University. Completed M.Tech(Electronic Instrumentation), National Institute of Tehnology Warangal. Completed PhD from Andra University.Completed Post Doc.Fellowship from Hoseo University South Korea.Currently Professor of Electronics and Communication Engineering and Director Industry Institute Interaction, Placements & Training Jawaharlal Nehru Technological University Kaknada Kakinada - 533 003, Andhra Pradesh, INDIA

Crosstalk Reduction Using Novel Bus Encoders in Coupled *RLC* Modeled VLSI Interconnects

G. Nagendra Babu, Brajesh Kumar Kaushik, and Anand Bulusu

The Microelectronics and VLSI Group, Department of Electronics and Computer Engineering, Indian Institute of Technology, Roorkee, Roorkee-247667, Uttarakhand, India
nagendra.babu.iitr@gmail.com, bkk23fec@iitr.ernet.in, anandfec@iitr.ernet.in

Abstract. Most of the encoding methods proposed in recent years have dealt with only *RC* modeled VLSI interconnects. For deep submicron (DSM) technologies, on-chip inductive effects has rapidly increased due to increasing clock frequency, decreasing signal rise times and increasing length of on-chip interconnects. This issue is an important concern for signal integrity and overall chip performance. Therefore, this research paper proposes an efficient Bus Encoder using Bus Inverting (BI) method. This method dramatically reduces both crosstalk and power dissipation in *RLC* modeled interconnects. The proposed encoder consumes significantly lower power which makes it suitable for the current high-speed low power VLSI interconnects. The proposed design demonstrates an overall reduction in power dissipation and crosstalk delay by 59.43% and 72.87%, respectively.

1 Introduction

The performance of a high-speed chip in deep sub-micron(DSM) technology is largely dependent on interconnects which connect different macro cells within a VLSI/ULSI chip [1]. With the ever-growing length of interconnect and on chip clock frequency, the effects of interconnects cannot be restricted to *RC* models. The importance of on-chip inductance is continuously increased with faster rise time, wider wires, and introduction of new materials for low resistance interconnects. It has become well accepted that interconnect delay dominates gate delay in current deep sub micrometer VLSI circuits. Inductance can increase the per unit length interconnect delay [2, 3] and cause ringing in the signal waveforms, which can adversely affect signal integrity. Furthermore, inductive effects in global interconnects are more severe due to lower resistance per unit length of line, as a result interconnect impedance becomes comparable to the resistive component. On the other hand, longer current return path has been achieved due to the presence of mutual inductive coupling between interconnects.

There are different methods for reduction of crosstalk such as repeater insertion, shielding line (V_{dd}/*GND*) insertion between two adjacent wires [4], optimal spacing between signal lines and lastly the most effective bus encoding method [5-10]. This paper uses bus invert method for reduction of power dissipation, crosstalk induced delay, propagation delay and chip size of encoder and decoder of *RLC* modeled interconnects.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 735–744.
springerlink.com

Here, the proposed method reduces two undesirable types of crosstalk *i.e.*, Type-0 and Type-1 couplings, which are worst case scenarios observed in *RLC* interconnects. Furthermore, the proposed design reduces power dissipation by reducing switching activity.

The organization of this paper is as follows. Section 2 describes crosstalk and power dissipation expression and their dependency on different parameters. The working of proposed method is discussed in section 3. Section 4 discusses the results obtained for encoder driving *RLC* modeled interconnects. Finally, section 5 draws important conclusions.

2 Power and Crosstalk in *RLC* Modeled Interconnects

The parasitic capacitance model of an interconnect consists of three parts, ground capacitance (C_G), fringe or sidewall capacitance to substrate (C_F) and coupling capacitance (C_C).

Coupling capacitance becomes dominant when adjacent wire tend to switch from 1 to 0 or 0 to 1, resulting in delay penalty which is called crosstalk delay. There are two important effects of noise on non-switching wires and increased delay on switching wires which occurs due to crosstalk.

All possible switching configurations can be classified in Type-0, Type-1, Type-2, Type-3 and Type-4 couplings depending on the value of miller coupling factor (MCF) as shown in Table 1.

Table 1. Classification of Crosstalk

Type-0	Type-1	Type-2	Type-3	Type-4
- - -	- -↑	- ↑ -	- ↑↓	↑↓↑
↑↑↑	- ↑↑	↑ - ↑	- ↓↑	↓↑↓
↓↓↓	↑ - -	↑ - ↓	↑↓ -	
	↑↑ -	↑↑↓	↓↑ -	
	- - ↓	↑↓↓		
	- ↓↓	- ↓ -		
	↓ - -	↓ - ↓		
	↓↓ -	↓ - ↑		
		↓↓↑		
		↓↑↑		

↑: switching from 0 to 1, ↓: switching from 1 to 0, - : no transition.

In *RC* model, the dominant factor is coupling capacitance, but for *RLC* model, mutual-inductance is considered to be dominant. Inductive coupling [11, 12] takes worst form when both of the interconnect lines which are adjacent have same transition (i.e., either from 0 to 1 (↑↑) or from 1 to 0 (↓↓)). In this case leftmost aggressor line induces magnetic field on victim line which tends to flow a current which is in opposite direction with respect to the original current [5,11]. So crosstalk occurs between two interconnects which is presently the major problem in DSM technology. Therefore, in *RLC*

model interconnects, when the lines are switching in same direction, then the worst case (↑↑↑ or ↓↓↓) [12] coupling occurs. Consequently, for *RC* modeled interconnects, worst case crosstalk delay occurs when adjacent lines are switched in opposite direction. However, worst case pattern in *RC* model is the best case for *RLC* model [5].

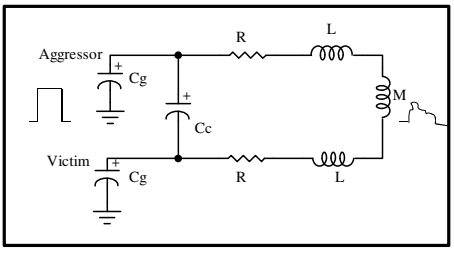


Fig. 1. *RLC* equivalent of an interconnect

Power dissipation is expressed as [7]:

$$P = \alpha. V_{dd}^2. f. C_L \quad (1)$$

where C_L is load capacitance, V_{dd} is supply voltage, f is the clock frequency and α is the average switching activity which lies between 0 and 1.

For reducing power dissipation in VLSI circuits, one or more factors such as V_{dd}, f, C_L and α must be minimized. Here, V_{dd} and f are assumed to be already optimized for low power. Therefore, dynamic power dissipation that is proportional to the number of signal transition can be reduced by optimizing switching activity. Symbols and terminologies used throughout this paper are as follows:

d(t): Bus value at the input of encoder.
D(t): Encoder output which is transmitted.
D(t-1): Encoder output which was latched up.
inv(t): Invert line at the input of encoder which is preset to '0'.
INV(t): Invert line for the encoded data sent at time *t*.
INV(t-1): Invert line for encoded data sent at time *t-1*.

3 Implementation of Proposed Design

This section presents proposed encoder and decoder using bus invert technique. The results obtained using proposed method is compared with previously published outcomes.

A. Proposed Encoder

The proposed encoder is a modified and improved version of Fan *et al*. [7] to reduce crosstalk, delay and power dissipation of *RLC* modeled interconnect instead of *RC* model. In this proposed method, data bus is divided into different clusters, where each cluster contains four data bits and one extra control bit. Basically, bus invert method [8, 9] utilizes an extra control bit i.e., *INV(t)* to differentiate the transmission of original

data and inverted data. In this method, if the number of transitions are more than half of the size of bus width, then original data is inverted and control line *(INV(t))* is set to 'high' whereas in other case original data is transmitted with *INV(t)* at logic 'low'.

The block diagram of proposed encoder is shown in Fig. 2 [10]. The 5-bit bus encoder architecture is composed of inverter, CNT0, CNT1_1, CNT1_2, 2-bit comparator, XOR stack and latch. CNT0 and CNT1 are called crosstalk modules which are used to count Type-0 and Type-1 couplings respectively. Two type-1 counters *viz.,* CNT1_1 and CNT1_2 are used which counts the number of type-1 couplings with original data and inverted data respectively. Brief description of the block diagram is as follows.

As soon as the data is transmitted (*d(t),inv(t)*) at the input of the encoder, it is passed through inverter whose output is ($\overline{d(t)}$, $\overline{inv(t)}$) as shown in Fig.2. Initially the value of *inv(t)* in the data to be transmitted is assumed to be at logic 'low'. Now, the original data (*d(t),inv(t)*) and previously latched data (*D(t-1), INV(t-1)*) are fed as inputs to CNT0 and CNT1_1 counters. The outputs of CNT0 (1 bit) and CNT1_1 (2 bits) are *N0* and K_1K_0 respectively. The inverted data ($\overline{d(t)}$, $\overline{inv(t)}$) and the data stored (*D(t-1), INV(t-1)*) are fed as inputs to CNT1_2 whose output (i.e., of 2 bits) is L_1L_0. The counts of two type-1 counters are compared in 2-bit comparator. The inputs of 2-bit comparator are K_1K_0 and L_1L_0 (which are having 2-bits) whereas the output of the comparator is *N1*(1-bit). Next, *N0* and *N1* are fed as inputs to an OR gate whose output is *INV(t)*. This *INV(t)* and the original data (*d(t),inv(t)*) are given as inputs to XOR stack.

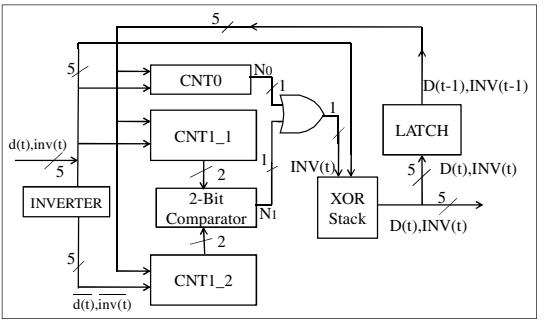


Fig. 2. Block diagram of 5-bit bus encoder

The output of XOR stack can be inverted data (if *INV(t)* is '1') or the original data (if *INV(t)* is '0'). The output of the XOR stack is the encoded data (*D(t), INV(t)*) which is finally fed to interconnects. This encoded data is stored in latch for one clock cycle (*D(t-1), INV(t-1)*), after which it is fed back for comparison with (*d(t), inv(t)*). Finally, at the receiving side, decoder retrieves the original data with the help of *INV(t)* line.

B. CNT0

CNT0 counts the number of type-0 couplings whose internal circuit diagram is shown in Fig.3. Type-0 coupling occurs if any of the three lines are transiting in the same

direction(i.e., ↑↑↑ or ↓↓↓). There are two type-0 couplings (one is of 'low to high' transition and the other is of 'high to low' transition) which can be merged to only one coupling. As 'low to high' transition (↑) and 'high to low' transition (↓) are detected separately, it is concluded that there is only one type-0 coupling. First, the design checks the occurrence of transition by using level-1 AND gates. The top five AND gates detect 'high to low' transition whereas bottom five will detect 'low to high' transition. A 'high' logic is present at the output if there is a transition (inversely a logic 'low' is present). The output signals from level-1 AND gates S_a,S_b,S_c,S_d,S_e ('high to low' transition) signals are grouped into three different combinations as ($S_aS_bS_c$, $S_bS_cS_d$, $S_cS_dS_e$). These signals are fed to three different AND gates to detect ↓↓↓ condition as shown in Fig.3. Similarly, the output signals of 'low to high' detector (S_f,S_g,S_h,S_i,S_j) are given to another set of AND gates to detect ↑↑↑ condition. Finally, all these signals are fed to OR gate, whose output (*N0*) becomes 'high' if any one of the combination satisfies the type-0 coupling condition.

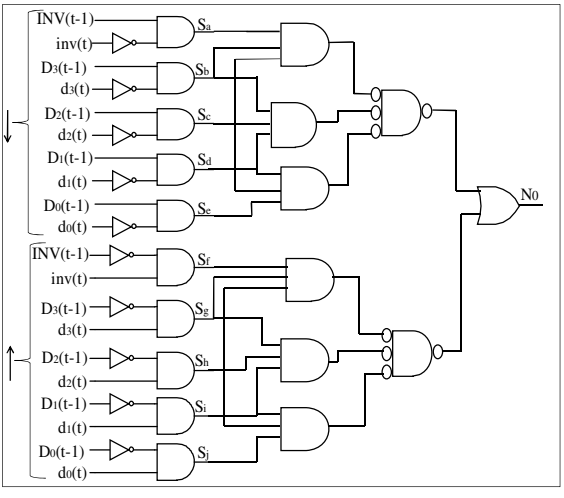


Fig. 3. Circuit diagram of CNT0

C. CNT1_1

Type-1 coupling occurs when one or two lines are having transition in the same direction while the rest (i.e., remaining two or the third one respectively) are idle. There are eight conditions of type-1 coupling which can be placed in two different groups with each group having four switching conditions that depends on high to low and low to high transitions. CNT1_1 counts the number of type-1 couplings with original data whose circuit diagram is shown in Fig.4. Level-1 AND gates of CNT1_1 detects the transitions as discussed in the above section. The outputs of level-1 AND gates are fed to OR gates. Type-1 coupling occurs when the first line is having transition and third line is idle and vice-versa, which clearly assures that these lines, must be inserted as inputs to an XOR gate to verify this condition. These five OR gate outputs are divided into 3 groups i.e., m_0 and m_2; m_1 and m_3; m_2 and m_4 which are fed to three different

XOR gates. The outputs of these three XOR gates should be added using a full adder. This method implements the full adder using two half adders and an OR gate and output of the full adder represents the number (as there are four in number two bits are sufficient to represent the count) of Type-1 couplings K_1K_0.

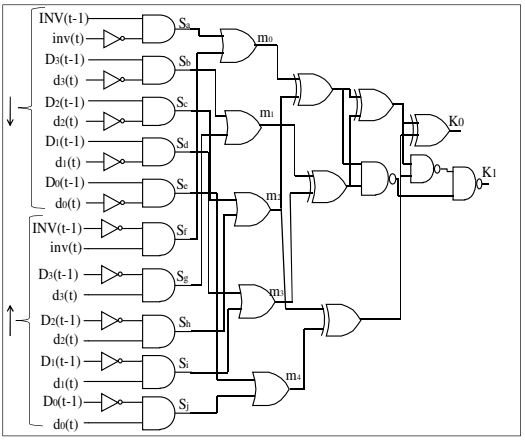


Fig. 4. Circuit diagram of CNT1_1

4 Results

The proposed encoder is simulated to find power dissipation and propagation delay of the bus codec. Simulation results were obtained using HSPICE in 180, 130, 90, 70 and 45nm technologies. Although internal diagrams of encoder has been shown for only 4-bit, the design can be extended to 8-bit and 16-bit using shielding method.

The power dissipation and crosstalk delay of the circuitry (encoder, interconnects and decoder) are obtained. Total power dissipation includes the power dissipated through encoder, interconnects and decoder (i.e. $P_{enc}+ P_{dec} + P_{interconnect}$). $P_{D,coded}$ is the dissipated dynamic power (which depends on switching activities of the data) after encoding of the data. In 180nm technology, when load capacitance is lesser than 0.1pF/bit, then coding method consumes more power than uncoded method. But as the load capacitance is increased *i.e.*, beyond 1.5pF/bit, then coded data consumes lesser power than uncoded data. Using proposed design for a load capacitance of 4pF/bit, a reduction in power dissipation ranging from 33.2% to 42.6% is achieved.

Table 2. Comparison of proposed model with fan *et al.* [7]

Component	Fan et al. [7]	Proposed Method
AND gate	4-input	2-input
6-bit adder	2	0
CNT0	2	1
XOR gate	18	8
Total no. of transistors	664	472

The crosstalk effect is reduced by inverting the original data which in turn also reduces switching activity. Here, the proposed method reduces both Type-0 and Type-1 couplings. By inverting all Type-0 and Type-1 cases, a reduction of 40% in switching activity is achieved which is substantially more as compared to Fan *et al*. [7] method. The proposed method greatly reduces the chip area by reduction in number of transistors as compared to Fan *et al.* [7] as revealed in Table 2. By reduction in chip area, the complexity of circuit is also reduced by more than 25%.

The encoder proposed by Fan *et al.* [7] uses four input AND gates whose input capacitance is very high and which in turn increases propagation delay. However, the proposed design uses only 2-input AND gates (except for six 3-input NAND gates) which effectively reduces encoder propagation delay. Moreover, Fan *et al.* [7] design uses two 6-bit adders which is comprised of four full adders and four half adders which makes the circuit more complex and occupies more area which results in higher power dissipation.

A. Total Power Reduction

Total power dissipated by the system includes the power dissipated by encoder, decoder and interconnects. The total power dissipated at 1GHz frequency in different technology nodes is shown in Table 3 and Fig.5. It has been observed that as feature size reduces, overall power dissipation also reduces.

Table 3. Power dissipation at different technology nodes

Technology (nm)	Power Dissipation (μW)		Power Saved (%)
	Proposed	Fan *et al* [7]	
45	288.32	775.72	62.83
70	775.84	1871.04	58.53
90	1422.45	3571.68	60.17
130	1764.75	5671.47	68.88
180	3521.97	8178.89	56.94

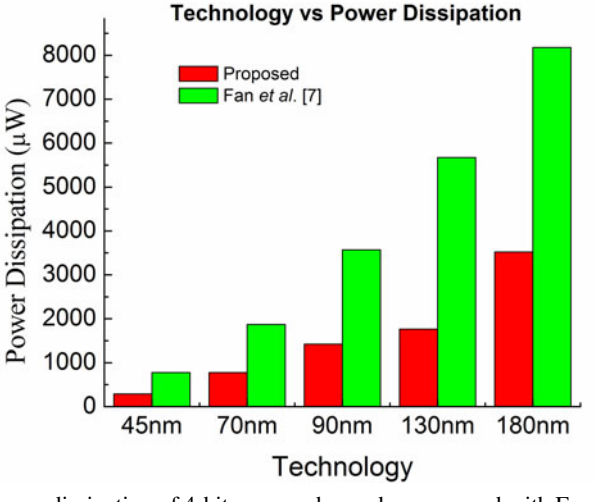


Fig. 5. Power dissipation of 4-bit proposed encoder compared with Fan *et al.* [7]

B. Total Propagation Delay Reduction

Propagation delay on a victim line increases with crosstalk. The encoder which is used to reduce crosstalk, also introduces some delay which is the overhead delay. Although there is a reduction in propagation delay with reduction of crosstalk, the overhead delay due to encoder should also be considered. The propagation delay introduced by proposed design is shown in Table 4 and Fig.6. Propagation delay is the delay occurred in signal as it propagates from encoder to decoder (i.e., encoder + interconnects + decoder). It is observed that as technology goes on shrinking, the overall delay decreases, which therefore acts as a trade off parameter for power dissipation. However, the overhead delay introduced by proposed design is lesser compared to other models. The results shown in Table 4 are obtained at 1GHz frequency.

Table 4. Worst case propagation delay in different technologies

Technology (nm)	Propagation Delay (ps)		Delay Reduction (%)
	Proposed	Fan *et al* [7]	
45	44.0563	190.52	76.88
70	102.458	340.35	69.90
90	102.512	334.52	69.36
130	120.5949	488.77	75.33
180	210.445	604.35	65.18

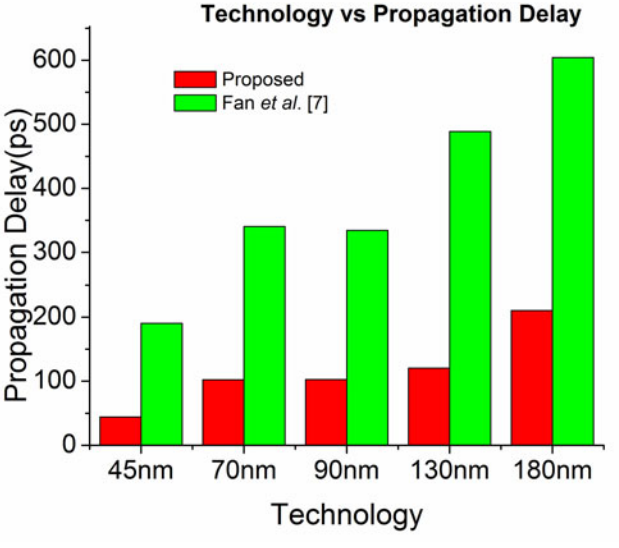


Fig. 6. Propagation Delay of 4-bit proposed encoder compared with Fan *et al.*[7]

5 Conclusion

The paper demonstrated a reduction in power dissipation, total propagation delay and crosstalk for *RLC* modeled interconnects by using low complexity encoder. This encoder consumes very less power as compared to the existing encoders which are used for crosstalk avoidance. The results show a reduction of 100% in Type-0 and 82.8% in Type-1 coupling. The proposed design demonstrated an overall reduction in power dissipation and crosstalk induced time delay by 59.43% and 72.87%, respectively. This codec system works upto 1GHz frequency without any problem. However, beyond this frequency there is a necessity to use a high speed latch in the feedback path.

Acknowledgment. The authors extend thanks to Special Manpower Development Project (SMDP-II), Ministry of Information Technology, Government of India. Without their support, research would have been a lot more painful experience than it already is.

References

1. Trevillyan, L., Kung, D., Puri, R., Reddy, L.N., Kazda, M.A.: An integration environment for technology closure of deep-submicron IC designs. IEEE Des. Test. Comput. 21(1), 14–22 (2004)
2. International Technology Roadmap for Semiconductors (2007)
3. Elgamel, M.A., Bayoumi, M.A.: Interconnect noise analysis and optimization in deep submicron technology. IEEE Circuits Syst. Mag. 3(4), 6–17 (2003)
4. He, L., Lepak, K.M.: Simultaneous Shield Insertion and Net ordering for Capacitive and Inductive Coupling Minimization. In: Int. Symp. Physical Design, pp. 55–60 (2000)
5. Shang-Wie, T., Yao-Wen, C., Jing-Yang, J.: RLC Coupling-Aware Simulation and On-Chip Bus Encoding for Delay Reduction. IEEE Transaction on Computer-Aided Design of Integrated Circuits and Systems 25(10) (2006)
6. Hirose, K., Yasuura, H.: A bus delay reduction technique considering crosstalk. In: Proc. Design Automation and Test Eur. (DATE), Paris, France, pp. 441–445 (2000)
7. Peng Fan, C., Hao Fang, C.: Efficient RC low-power bus encoding methods for crosstalk reduction. Integration, the VLSI Journal 44(1), 75–86 (2011)
8. Victor, B., Keutzer, K.: Bus encoding to prevent crosstalk delay. In: Proc. Int. Conf. on Computer-Aided Design, pp. 57–63 (2001)
9. Stan, M.R., Burleson, W.P.: Bus-Invert Coding for Low-Power I/O. IEEE Trans. VLSI Syst. 3, 49–58 (2005)
10. Nagendra babu, G., Agarwal, D., Kaushik, B.K., Manhas, S.K.: Power and Crosstalk Reduction using Bus Encoding Technique for RLC Modeled VLSI Interconnect. In: 2nd International Workshop on VLSI, pp. 424–434 (2011)
11. Ismail, Y.I.: On-Chip Inductance Cons and Pros. IEEE Trans. on VLSI Syst, 685–694 (2002)

12. Chowdhury, M.H., Ismail, Y.I., Kashyap, C.V., Krauter, B.L.: Performance Analysis of Deep Sub micron VLSI Circuits in the Presence of Self and Mutual Inductance. In: IEEE Int. Symp. Circuits Syst., pp. 197–200 (2002)
13. Baek, K.H., Kim, K.W., Kang, S.M.: A Low Energy Encoding Technique for Reduction of Coupling Effects in SOC Interconnects. In: Proc. 43rd IEEE Midwest Symp. Circuits Syst., pp. 80–83 (2000)
14. Weste, N., Eshraghian, K.: Principles of CMOS VLSI Design, A Systems Perspective. Addison-Wesley Publishing Company, Reading (1988)
15. Rabaey, J.M., Anantha, C., Borivoje, N.: Digital integrated Circuits: A Design Perspective, 2nd edn. Prentice Hall Publication (2003)

Event Triggering Mechanism on a Time Base: A Novel Approach for Sporadic as well as Periodic Task Scheduling

Ramesh Babu Nimmatoori[1], A. Vinay Babu[2], and C. Srilatha[3]

[1] Research Scholar, Department of CSE, JNTUH, Hyderabad, India-500085
[2] Professor, Department of CSE, JNTUH, Hyderabad, India-500085
[3] Assoc Prof. and Head, Department of ECE, ASTRA, Hyderabad, India-500008
rntoori@yahoo.com, avb1222@gmail.com, deepuaurora@yahoo.com

Abstract. A novel concept involving the combination of the two individual time triggered and event triggered methodologies is presented aiming in gaining the benefits of the respective mechanisms and attaining high system level performance. The major advantage of the work presented is the best utilization of the system resources available with gained flexibility. A detailed explanation of the proposed approach is presented. Simulation studies were conducted for the approach validation.

Keywords: Worst case latency, Message transmission time, Real Time Distributed system, time trigger, event trigger, periodic task, sporadic task.

1 Introduction

Real Time Distributed Systems are getting complicated and diverse day by day. Depending upon the application, the choice of either the time triggered mechanism or event triggered mechanism was opted. A Real Time Distributed System employing the time triggered architecture demands an initialization and synchronization stage. The worst case latency estimation is a key issue and to be estimated under all possible operating conditions. With all these efforts and study, the timing behavior thus turns to be predictable. On the other hand, as the event triggered architecture does not require the initialization phase because of the unsynchronized nodes, the timing behavior is indeterminist. Also event triggered systems need exhaustive simulation study compared to that of the RTDS employing time triggered methodology. As the event triggered RTDS does not support any temporal encapsulation, the scalability feature is very much poor. But the resource utilization fashion is well available in such systems. As today's RTDS are becoming more and more heterogeneous, there exists a demand to cater the needs of both time and event triggered situations simultaneously. This situation arises as some of the modules may be time triggered while some of them may be event triggered components. After the compilation of the two methodologies and the detailed simulation study, the implementation benefits were drawn out. An interesting research corner with respect to gaining the benefits from both the methodologies if combined came into issue. The optimistic solution for such type of heterogeneous RTDS is the combinational implementation of both the methodologies which would have a significant impact on the system.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 745–754.
springerlink.com

2 Previous Work

There has been an exhaustive research work carried out on the implementation of RTDS whether to opt the time triggered or the event triggered mechanism [1, 2 and 3]. Many parameters like flexibility, resource utilization, scalability etc were studied comparatively. These works concluded that the choice of the paradigm depends on the requirements of the application. Today's fast growing heterogeneous RTDS and future application systems demand for the coherent presence of both time and event triggered tasks [4] and hence there exists a need for their mutual interaction over the communication network medium. When both the tasks share the same node, the architectural support is to be in accordance. Such a heterogeneous distributed system employs the interaction and exchange of both static and dynamic messages ie. Periodic and sporadic. This paper presents an introduction to heterogeneous distributed systems and the significance of the sporadic data. Also a new combinational approach is introduced. The applications running on such mechanisms are very difficult to analyze. Because of the hierarchical nature of the modules, multiple execution interferences (conflicts/message collisions) occur. They have to be carefully accounted during the timing analysis that determines the worst-case latency of the system. Along with this, the message delays are to be considered. The timing analysis is further complicated by the respective characteristics of the communication protocol employing both the time triggered and event triggered paradigms. Communication protocols employing the combinational architectures were into research [5, 6 and 7]. The major drawback of them was that the basic benefits of time triggered architecture were lost. Our research work focuses on preserving the individual advantages and overcoming the disadvantages. Hence, In order to meet the design challenges of such heterogeneous RTDS, an adequate environment is to be developed to effectively support cost-efficient and high performance.

3 Significance of Sporadic Messages

The sporadic messages acquire the periodic frame which is not being used by any other periodic message in a time triggered system. With respect to the bandwidth utilization, they are well suited for periodic messages only. The sporadic messages are better handled by an event triggered system. But, practical applications demand the information of both periodic and sporadic types. Hence, current day protocols require a suitable scheme that combines the two.

Handling Low Priority Messages

Low priority sporadic messages are to be handled by the time triggered RTDS by employing appropriate scheduling methodology and also ensure that the respective deadline which are soft in nature, are met. To meet this requirement, a predefined amount of slack is scheduled by each node in the system which is being dedicated for the sporadic messages. The above figure depicts the allocation of sporadic messages as part of the data frame. The black bars show the sporadic messages.

Fig. 1. Cluster with sporadic messages

As the frame size is fixed and also the periodic message size is predefined, the frame size available for the sporadic messages is fixed. The address for these messages is to be included as a part of the message itself while for that of the periodic messages it is implicit because of the statistical scheduling. With this approach, as the slack is prefixed, the node flexibility is limited. In other words, one node cannot make use of the free slack of the other. This leaves aside the prioritization on sporadic data because they are handled by each of the node internally. To overcome this scenario, the periodic messages can be scheduled first so that there exist no slack in the individual nodes. For this, each of the communication round is equally divided into two portions in order to cater periodic and sporadic messages respectively. The system nodes employ the TDMA scheme during the first part of the communication round for the periodic data transfer while the second part for the sporadic data transfers. The system nodes employ some other access methodology for the communication medium access. But whenever there exist any free slots in the TDMA communication round, it will be occupied by the sporadic data.

Handling High Priority Messages

In some of the critical applications, these may be instances where sporadic messages can be even more of higher priority than that of the periodic messages. In other words, whenever any periodic message is being executed a higher priority sporadic message will interrupt it and after the successful execution the control comes back to the periodic message which employs the TDMA access. It is also assumed that the higher priority interrupting message length is fixed. The current approach of handling them is to employ an interrupt process which is initiated by a triggering clock pulse to interrupt the TDMA access and allow the higher priority sporadic messages. So whenever an interrupt is caused to be generated by a trigger pulse it is thus assumed to immediately transfer the sporadic message first and the periodic message later. Clock and node synchronization is essential in this case. Nodes need to have all the information regarding the resume time after the higher priority sporadic message execution, which node to resume, time frame etc. this mechanism leads to architectural complexities in the system design. With the occurrence of interrupts, the timing deterministic characteristics of the system would be lost in some cases. One precaution to be taken in this approach is to ensure that these pulses should not be raised frequently as they cause conflicts among the sporadic messages.

4 Implementation Scheme

This chapter mainly focuses on the combinational concept for ensuring the event triggered data to be run on a time base, which would easily handle both periodic and

sporadic messages efficiently. This is because, the time triggered methodology exclusively handles the periodic messages only while that of the event triggered handles sporadic data. In order to cater the needs of the current day heterogeneous RTDS, this combinational concept is very much beneficial.

The proposed approach is to ensure the consideration of high priority sporadic messages which can even occur more frequently in a RTDS. This methodology employs a "gap slot" concept wherein the above mentioned sporadic message collisions are overcome. Figure 2 shows the normal TDMA scheme.

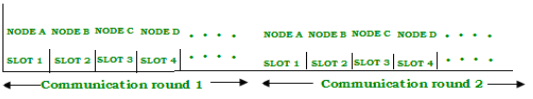


Fig. 2. TDMA scheme for two rounds of communication

The gap slot concept is depicted in fig.*3* below which shows the gaps between each node slots of the fig.*2*. This approach utilizes these gaps to arbitrate among the system nodes which are dedicated to send higher priority sporadic messages. The gap length is proportional to the number of gaps and the number of system nodes that allow sporadic message flow.

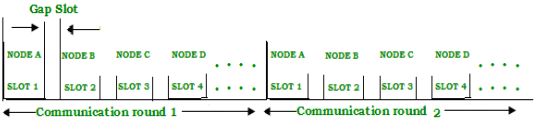


Fig. 3. TDMA scheme with gap slots for two rounds of communication

The fig 4 clearly shows how a higher priority sporadic message is being sent by anode (node C).

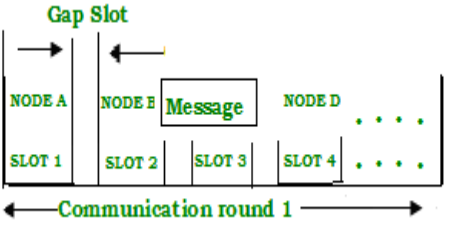


Fig. 4. TDMA scheme with gap slot being utilized by a higher priority sporadic message in a single round of communication

Here we present an implementation scheme for this architecture based on the gap slot concept. In this approach the sporadic messages (event triggered data) is being sent on a time triggered channel employing the TDMA scheme. As assumed in the previous chapters, there would be no preemption of the communication for sending

higher priority messages. But the lower priority sporadic messages are restricted in case of short execution times. Each of the system nodes will allot a gap slot for the sporadic data as per the statistical scheduling. The address for the sporadic messages is included in the message itself. With this it is possible to get better bandwidth. The major disadvantage of gap slot approach is that only the respective node can access the gap slot. This poses a limit on the communication bandwidth.

5 Proposed Approach

Here we propose a new approach for the sporadic message transmission namely the empty slot access. The gap slot concept is slightly modified in such a way that the gaps put together will occur after the allocation of all the system node slots. The ultimate aim of this proposed approach is to obtain short access time and enhanced performance in message transmission among the nodes. A detailed study has been made on how the system nodes will access the gap slots on mutual basis. Sporadic messages are queued internal to a node on priority basis. Periodic messages are transmitted directly over the communication network medium while that of the sporadic messages need some additional information along with the data. As the sporadic data is not static, it requires the following:

1. start bit
2. stop bit
3. message length
4. message ID (message address)

We first tried to define the empty slots of the TDMA communication round. An empty slot is the same as that of the gap slot defined in the previous section but with a difference that the gap slot occurs in between each node slots. After calculating the number of nodes and the number of gap slots the individual length of the gap slots is calculated. Here, we have made a beneficial change in such a way that all the gap slots are put at the end of the communication round following the node slots. The major advantage of this approach is that the bandwidth is open to all the system nodes wherein gap slot can be used by any of the node. This can accommodate large amount of event triggered data.

In order to study the efficiency and performance estimation of a RTDS handling sporadic and periodic data on equal importance basis employing the proposed combinational architecture, we have implemented the following two approaches:

1. defining the empty slot globally
2. defining the empty slot locally

Defining the Empty Slot Globally

Here one of the system node, call the master node, will define the empty slot globally in such a way that all the other nodes can send the respective messages as per the schedule made. The master node prioritizes the requests and schedules which sporadic messages to send and at which time. Thus the master node will include this information as a part of its own message. The following figure depicts the mechanism.

Fig. 5. Concept of empty slots

Where ES = empty slot
SPM = sporadic message

Other than the overhead calculated as per the constraints in chapter 5, an extra overhead for the sporadic data schedule is encountered in this method. One of the major remarkable factor of this approach is that the schedule transmission over the communication network medium. The major advantage of this type of approach is that a single node controls the entire scheduling process. At the same time, it alone can cause single point of failures.

Defining the Empty Slot Locally

To overcome the disadvantage of the above presented approach, we have further investigated an extension of the gap slot scheme in such a way that the empty slots are defined locally. If you closely observe the fig 3 and this fig, both seem to be the same except the nomenclature. In fig 3, a fixed node has a fixed gap slot to send the message while in this approach, each of the system node schedules itself depending on the message it receives in such a way that which empty slot to access. This is designed in a fault tolerant aspect. Thus each of the nodes defines a local schedule to access an empty slot. This aims at less overhead because of the presence of sporadic messages and less delay in communication medium access. The performance efficiency of sporadic message transfer using this approach depends on the number of empty slots. The scheduling algorithm employed by the system nodes is the earlier request first, in other words, the first come first serve request. Which ever node has a first request that will be given higher priority. With this approach, if the number of system nodes is greater than the number of empty frames, by default some of the nodes will not get access to the empty slots. The following figure depicts the local empty slot mechanism.

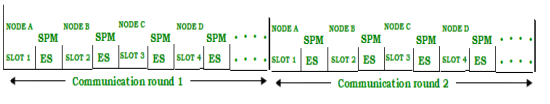


Fig. 6. Local empty slot mechanism

With this approach, the access time of the event triggered data is minimized because of the availability of empty slot alternatively in the TDMA communication round. These distributed alternate empty slots are spread across the entire TDMA communication rounds. Using this approach, in case of a node failure to get a message request for an empty slot, it will get the schedule of the overall positions of the other empty slots throughout the communication round. Thus, in the further communication rounds, the node can easily reintegrate its sporadic message into its respective empty slot for transferring over the communication channel. The notable remark of this approach is that for each TDMA communication round, each of the

system node is allowed to requeue its message requests. Whenever it is the turn of any node to send its sporadic message (event triggered data), say node A, it is removed from the queue irrespective of the validation whether it is being assigned an empty slot or not. If in the case an empty slot is not assigned for node A, it should once again request for the empty frame in its local queue. Hence, after one successful TDMA communication round, the global queue is accessed and node A gets its request processed and can gain access to the empty slot in the second TDMA communication round. With this it can be clearly noticed that the event triggered data works in coordination with the global queue while the local queue of a node works on the priority basis of its sporadic messages.

Resource Optimization

The above introduced queuing methodology will be further best useful if a node sends its message priority is sent along the message request to access an empty slot. But this introduces an additional overhead as message priority is also sent on the communication network medium. As a number of message parameters are to be set before the simulation run, the number of message priorities can be limited on trail basis. This limitation in turn limits the number of physical request bits required to transfer both the message request and its priority. If the number of empty slots are very less per TDMA communication round and in the case wherein all the system nodes want to send the respective sporadic messages the lower priority nodes may not get chance to access the empty slots. They need to again wait for the second communication round subject to the condition that all the higher priority nodes are finished with their message transmission. In worst cases, some of the lower priority nodes may not get this chance at all. Hence, they starve for the empty slots. This starvation phenomenon can be overcome with the round robin scheduling algorithms. A Special bit can be assigned in the data frame to indicate the presence of a sporadic message to be transferred by a node. When this special bit is zero, it means the node does not have any sporadic data to send and hence the empty slot can be utilized by some other node which needs to transfer the sporadic data. This has an advantage of best utilization of the communication network medium access as no empty slot is being allocated to a node which has no sporadic data to transfer.

6 Validation

In this section we present the simulation results obtained for the presented approach using the SICStus Prolog tool. Here we have studied the two major parameters namely the wait time and the number of missed messages against the system load. The localization of empty slots being advantageous over the global empty slots more focus was given to the former approach. The goal of this research work is to obtain the flexibility in handling event triggered data over the time triggered protocol to gain high performance of the overall RTDS. To achieve the set goal, the following were the assumptions made for the simulation study:

1. Considerable approach - localization of empty slots
2. All the time triggered slots are of equal length

3. All the empty slots are of equal length
4. The length of the time triggered slot and that of the empty slot is not same

The following figure depicts the simulation traces of the event triggered load against the message wait time in the queue. The straight line gives the simulation trace obtained with the time triggered approach while the dotted line gives the simulation trace obtained with the proposed combinational concept.

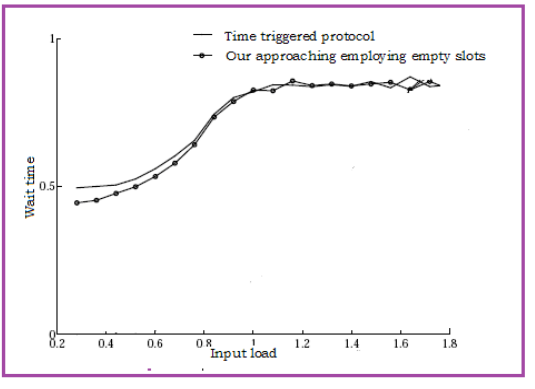


Fig. 7. Comparison between the time triggered and combinational approach

It can be clearly observed that the proposed combinational approach also maintains almost the same wait time for sporadic messages. For low loads, the wait time for the sporadic messages still lowers. In other words, this research work benefited in attaining high speed sporadic message transfers.

The following figure shows the simulation traces of the event triggered load against the number of missed messages.

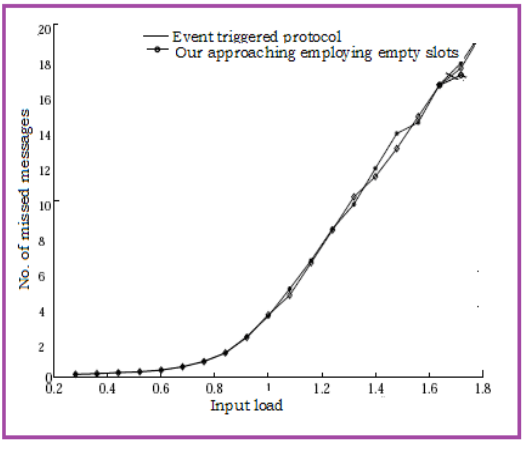


Fig. 8. Comparison between the event triggered and combinational approach

The number of missed messages is comparatively low for this presented approach. At low loads, the number of missed messages is negligible. In other words, at low loads system is highly efficient in terms of its speed and performance.

7 Results

The performance estimation of the presented combinational distributed methodology is analyzed with respect three important parameters. They are:

1. Worst case latency
2. Message transmission time
3. Maximum overhead

The following table shows the research work results as a comparison against the two classic approaches being presented in chapter 4 and 5 respectively.

Table 1. Comparative analysis

Implementation scheme	Worst case latency	Message transmission time	Maximum over head
Time triggered approach	D_{ti}	$EM_{ti,tj}$	0
Event triggered Approach	ML	$D_{ti}/2$	(n + SP)oh
The proposed approach presented in chapter 6	D_{ti}	$D_{ti}/2$	*n*

1. Worst case latency

The results clearly show that the proposed approach gains the advantage of the time triggered approach with respect to the execution speed as the worst case latency of both the approaches is the deadline assumed during the initial system designing phase. The worst case latency for a purely event triggered system is the maximum message length the system nodes can deliver.

2. Message transmission time

As the approach employs alternate empty slot mechanism, the message transmission time has reduced to half the time compared to the traditional time triggered approach. This is because that the slack cannot be altered in case of the time triggered approach, each node gets exactly one fixed slot per TDMA communication round. On the other hand, with the presented combinational approach, each node may even access more than one empty slot depending upon its priority levels. This drastically increased the message transmission time, amount of data transfer and hence the speed of the system.

3. Maximum overhead

Overhead is exclusively with respect to the presence of sporadic data in a RTDS. Because of the predefined slack, ant time triggered system does not suffer from overhead. But in case of an event triggered system the overall overhead depends on the number of nodes, number of sporadic messages and the overhead per message frame. Put together, for high loads it introduces considerable amount of overhead. This disadvantage is overcome with the proposed combinational concept wherein the maximum overhead does not cross the available number of system nodes.

8 Conclusions

With the above presented approach, it can be concluded that it is possible to design a combinational distributed concept having the timing determinism feature of the time triggered methodology along with the flexibility and resource utilization advantage of the event triggered methodology. It is noticeable that both the time and event triggered data are handled equally.

9 Advantages

The combinational concept thus analyzed results in the following advantages:

1. The gap between the time triggered and the event triggered system is bridged with the approach presented.
2. Improvement with respect to fault tolerance.

References

[1] Audsley, N., Tindell, K.A., et al.: The End of Line for Static Cyclic Scheduling? In: 5th Euromicro Workshop on Real-Time Systems (1993)
[2] Xu, J., Parnas, D.L.: On satisfying timing constraints in hard-real-time systems. IEEE Transactions on Software Engineering 19(1) (1993)
[3] Douglas Locke, C.: Software Architecture for Hard-Real Time Applications: Cyclic Executives vs. Fixed Priority Executives. Journal of Real-Time Systems 4, 37–53 (1992)
[4] Koopman, P.: Critical Embedded Automotive Networks. IEEE Micro 22(4), 14–18 (2002)
[5] Leen, G., Heffernan, D.: TTCAN: A New Time-Triggered Controller Area Network. Microprocessors and Microsystems 26(2), 77–94 (2002)
[6] Führer, T., Müller, B., Dieterle, W., Hartwich, F., Hugel, R., Walther, M.: Time Triggered Communication on CAN (Time Triggered CAN - TTCAN). In: 7th International CAN in Automation Conference, ICC, Amsterdam, pp. 92–98 (2000)
[7] Belschner, R., Berwanger, J., Ebner, C., Eisele, H., Fluhrer, S., Forest, T., Führer, T., Hartwich, F., Hedenetz, B., Hugel, R., Knapp, A., Krammer, J., Millsap, A., Müller, B., Peller, M., Schedl, A.: FlexRay – Requirements Specification. FlexRay Consortium, Internet: Version 2.0.2 (April 2002), http://www.flexray.com

A High Level Approach to Web Content Verification

Liliana Alexandre[1] and Jorge Coelho[2]

[1] Lusitânia - Companhia de Seguros, S.A.
liliana.alexandre@lusitania.pt
[2] School of Engineering of the Polytechnic of Porto (ISEP) &
Artificial Intelligence and Computer Science Laboratory of the
University of Porto (LIACC)
jmn@isep.ipp.pt

Abstract. In this paper we present a tool for visually imposing constraints over the content of XML-based webpages and automatically repair such webpages in case they don't comply with the imposed constraints. The tool is based in the XCentric programming language and relies on a highly declarative model.

1 Introduction

VeriFLog [9] is an extension of the XCentric language [11] for semantic verification of XML-based content. It relies on the unification with terms with flexible arity symbols and sequence variables which enables a compact description of constraints. It also adds builtins to enhance the development of programs in the content verification domain. The main drawback of VeriFLog is that the user needs to have at least some basic knowledge of Logic Programming in order to use it. The tool presented here enhances VeriFlog by capturing the core features and adding new ones in a user-friendly visual approach which reduces the need of previous knowledge of Logic Programming. The main application of this tool is to verify content on collaborative websites such as Wikipedia [13].

The remaining of this paper is organized as follows, in section 2 we explain briefly the main concepts behind the XCentric language and the VeriFLog tool. Then, in section 3 we show how to compose rules for verifying XML-based webpages using our visual approach. In section 4 we present the related work and finally in section 5 we conclude and present future work.

We assume that the reader is familiar with Logic Programming [17].

2 Verifying XML Content

Here we explain how to verify content in webpages by using the XCentric language [11] and VeriFLog [9].

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 755–764.
springerlink.com

2.1 XCentric

XCentric extends Prolog with terms with flexible arity symbols and sequence variables. This approach enables a highly declarative model for querying content in webpages. Constraints of the form $t_1 = * = t_2$ are solved by a non-standard unification that calculates the corresponding minimal complete set of unifiers. Details about the implementation of this non-standard unification can be found in [8]. In XCentric an XML document is translated to a term with flexible arity function symbol. This term has a main functor (the root tag) and zero or more arguments. Although XCentric translates attributes to a list of pairs, we will omit them for the sake of simplicity. Consider the simple XML file presented bellow:

```
<addressbook>
     <record>
          <name>John</name>
          <address>New York</address>
          <email>john.ny@mailserver.com</email>
     </record>
     ...
</addressbook>
```

The equivalent term is:

```
addressbook(record(name('John'),
                   address('New York'),
                   email('john.ny@mailserver.com'))
                   ,...)
```

If we want to get the names of the people living in New York and assuming that the document is stored in variable *Doc* we can simply solve the following constraint:

$$\text{Doc} = * = \text{addressbook}(\_, \mathit{record}(\text{name(N)}, \text{address('New York')}, \_), \_).$$

All the solutions can then be found by backtracking (in variable N).

Note that '_' is an unnamed sequence variable which unifies with any sequence. So, no matter how many records the address book has, we can describe our constraint in a very compact way by focusing on the ones that matter for our purposes. The details of the language and several illustrating examples can be found in [11]. Although the operator $= * =$ supports variables in both sides we implemented a version which supports variables only on the right-hand side (operator $=\sim=$). This is enough for processing and querying documents (which don't have any variable inside) and increases performance. So, in the previous example, since *Doc* is an XML document without any variables, the operator $=\sim=$ could be used, giving the same results. In the tool we describe in this paper we only use operator $=\sim=$.

2.2 VeriFLog

In [9] and [10] XCentric was extended with several features to enable specific applications to verify, query and filter content in webpages that include:

- Definition of simple rules for website verification and filtering namely, replacing, deleting and blocking content.
- Use of types for static and dynamic verification of rules.
- Consistency checking between rules (one rule cannot violate another rule).

Let's present one simple example which illustrates how a *delete* rule can be implemented in VeriFLog.

Example 1. Given a wiki webpage in an XML document stored in variable *Wiki1*, deleting all the references in the text which do not occur in the bibliography section of that given wiki webpage is done by the following code:

```
delete(ref(R),Wiki1,Wiki2,[not(deep(bibentry([(number,R)],_),Wiki1))]).
```

So, if we have the following XML stored in variable *Wiki1*:

```
<?xml version="1.0" encoding="utf-8"?>
<WikiArticle>
....
<Content> XCentric <ref>3</ref> is an extension of Prolog with
unification of terms of flexible arity which enables a simpler
and high level querying and processing of XML data.
</Content>
...
<References>
  <bibentry number = "1">Jorge Coelho and Mario Florido.
  XCentric: Logic Programming for XML Processing. 9th ACM
  International Workshop on Web Information and Data
  Management. ACM Press, 2007.</bibentry>
   <bibentry number = "2">SWI-Prolog,
          http://www.swi-prolog.org/</bibentry>
  </References>
</WikiArticle>
```

By applying the delete rule and since a reference with number 3 is not available in the references at the bottom of the page (attribute *number* of element *bibentry*) it will result in a new XML document in variable *Wiki2* where the element $< ref > 3 < /ref >$ was deleted.

The *replace* and *failure* rules work in an analogous way. The type system allows checking the content against schemas and the consistency checking verifies if one rule is not in violation of another rule, for example, when one rule adds some content which is forbidden by another rule.

3 Visual Editor of Rules

With the tool we describe here a user can select the XML Schema (XSD) [23], describe constraints over documents complying with the given schema and then apply these constraints to instances of that schema. In case the schema is not available, the user can select an XML document and the application will infer the corresponding XSD. It is possible to select sub-trees of the document and apply constraints to its content, such as string manipulation, negation, emptiness and URL checking. It is also possible to introduce constraints manually in order to search elements at arbitrary deep and apply complex constraints to these in a highly declarative and compact syntax. Details and examples are presented next.

3.1 Implementation

This tool is implemented in C# [21] and SWI-Prolog [24]. For the communication between C# and SWI-Prolog we use a third-party library named Swi-cs-pl [22]. In Figure 1 we present the main interface of our application. Here the user can choose between two types of file for the input, an XSD or an XML instance. In the case the user chooses the XML instance, the application infers the related XSD. The user can also choose one of two ways for applying the rules, applying to a unique file or to a directory of files. The idea for choosing a directory is that the user can verify the constraints to a set of files conforming to a given XSD. Note that the XSD was loaded and presented in the left tree view. The user can now proceed by selecting sub-trees and applying rules to these. When applying the constraints, the application first checks if the input file complies with the related XSD.

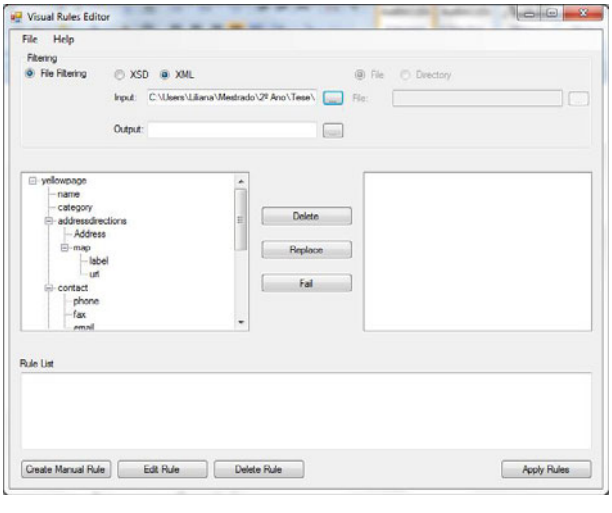


Fig. 1. Main interface

3.2 Examples

We now present some illustrating examples. For these examples, we use a wiki that stores yellow pages where anyone can contribute and which is available at [25].

Example 2 (Deleting content). In these wiki-based yellow pages there is a section where a user can insert a URL with a link to a map showing the location of his/her business, this URL is in a subtree like the one presented next.

```
...
<map>
  <label> ... </label>
  <url> ... </url>
</map>
...
```

The user may, for example, write the content of the *label* element but forget to include the content of the *url* element. We may argue that this doesn't make sense and impose a rule that checks the content and removes the subtree *map* whenever the *url* element is empty. In Figure 2 we show how this is done. We selected an XML file whose XSD was inferred and presented in the left side. There, the user can select the element to which he wants to apply the constraint. For this example we select the element *map* (the one we want to delete) and click the *Delete* button to open the rule definition window. There, *Element content* dropdown is filled with the elements contained in the subtree of *map*, we can choose any of these and define constraints over their content. These constraints consist in optionally picking the "NOT" checkbox and choosing one of the "Contains", "Contains valid URL" or "NULL". In this case we choose the *url* element and pick the "NULL" checkbox. After clicking the *Apply* button the rule is added to the rule list and the right tree view is loaded with the new version of the XML file. As shown in Figure 3 the *map* element does not appear anymore in the final document. The generated rule is presented next:

```
delete(map(Map), YP1, YP2,[deep(seq(url([],U),empty), Map),(U=~=empty)]).
```

Here, *YP1* stores the initial XML document and *YP2* stores the new XML document after applying the constraint.

Example 3 (Replacing content). Given the same wiki webpage presented in the example above, we want to validate if the prices are not missing. If they are missing we want to replace the null content of these with a warning message such as "Prices Unavailable". Using the visual rule editor, one can select the element *prices* and click the *Replace* button to define this rule. Here we click the *NULL* checkbox to verify if the content of *prices* is empty. After clicking the *Apply* button the *Rule List* is updated with this rule and the new XML (on the right side) is updated.The generated rule is presented next:

```
replace(prices([],P),prices([],'Prices Unavailable'),YP1,YP2,
        [deep(seq(prices([],P),empty), YP1),(P=~=empty)]).
```

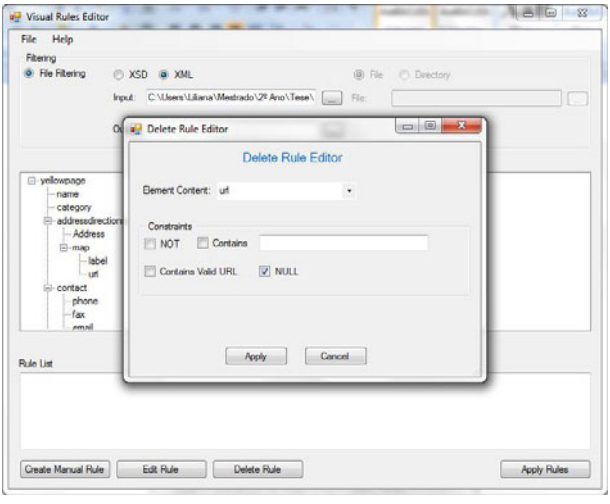


Fig. 2. Applying a delete

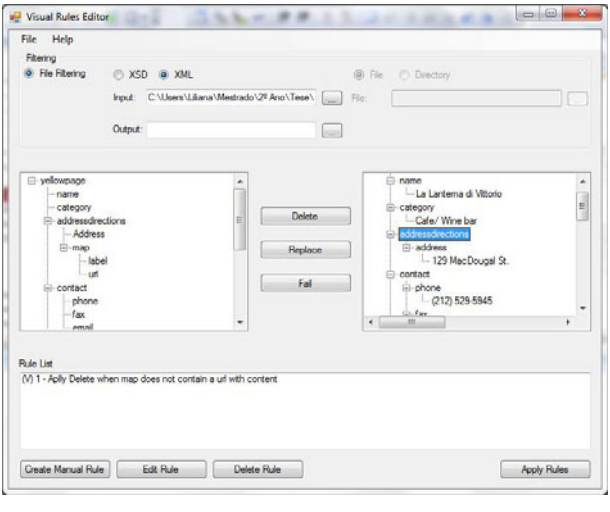


Fig. 3. Result of delete

Here, *P* stores the content of element *prices* and *YP1* the initial XML document and *YP2* the new XML document after applying the constraint.

Note that rules are being added to the *Rule List* at the bottom of the main window interface. These rules can be all applied to an XML document we choose. Given the following XML file (stored in variable *YP1*):

```
<yellowPage>
...
   <addressdirections>
      <address>
         129 MacDougal St.
      </address>
      <Map>
         <label>Map it</label>
         <url></url>
      </Map>
   </addressDirections>
...
   <prices></prices>
...
</yellowPage>
```

By applying the two rules presented in the examples above the new XML document stored in *YP2* variable has the same document but with:

```
...
   <prices>Prices Unavailable</prices>
...
```

Example 4 (Disapproving webpages). An error found in a document can be seen as so severe that it is better to stop the page processing and present an error message. This could be useful if, for example, this tool was automatically integrated in a website such as Wikipedia to automatically verify errors in content of submitted webpages.

Let's consider that an invalid email is a severe error. We will implement a simple verification by checking if the email contains an @. If it does not contain an @ we will just present an error message and will not process the XML document. We do this by selecting the *Fail* button as presented in Figure 4. The generated rule is presented next:

```
failure(YP1,[deep(seq(email([],E),empty),YP1),
             not(sub_string(E,_,_,_,'@'))],'Valid email not found!').
```

Here, *YP1* contains the input XML file and the variable *E* variable contains the email content to verify. In case of error the message "Valid email not found" is shown to the user.

Example 5 (Describing rules manually). Using the basic rules in the interface windows we may be unable to verify every aspect we need. Thus, the editor gives the possibility of manually editing rules in order to use all the power of Prolog and XCentric. Let's suppose we want to delete all phone numbers which length is not equal to 10. This can be done by clicking the button *Edit Manual Rule* and inserting the following rule:

```
delete(phone([],P),YP1,YP2,[deep(seq(phone([],P),empty),YP1),
                            (name(P,L),length(L,LP),LP=\=10)])
```

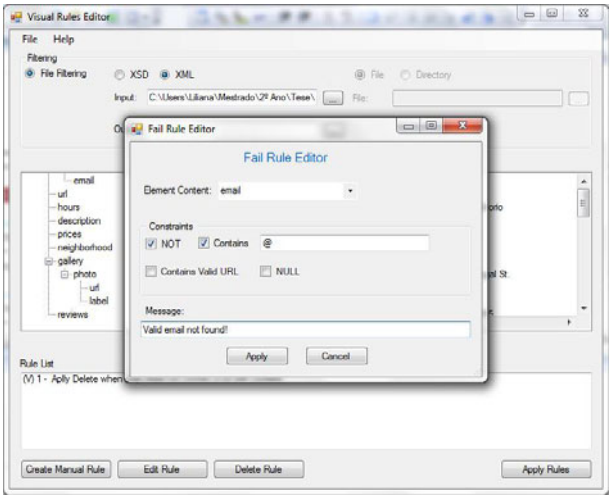


Fig. 4. Disapproving webpages

Also, note that it is possible to delete previously defined rules by selecting them and clicking in "Delete Rule" updating the XML in the right tree view with a version where the deleted rule was not applied. It is also possible to save these rules for reusing in the future.

4 Related Work

The tool presented here is a visual extension with new features to our previous work presented in [9] and [10]. A preliminar version of this work was presented in [2] and [3]. The new version presented here contains several improvements and bug fixes which make the tool more capable and easy to use. In [4] the authors presented a rewriting-based framework that uses simulation [15] in order to query terms. In [5], the authors present a semi-automatic methodology for repairing faulty websites by applying a set of concepts from Integrity Constraint [20]. In [12] the author proposed the use of a simple pattern-matching-based language and its translation to Prolog as a framework for website verification. In [14] logic was proposed as the rule language for semantic verification. There the authors provide a mean for introducing rules in a graphical format. In [16] the author proposed an algorithm for website verification similar to [7] in expressiveness. The idea was to extend sequence and non-sequence variable pattern matching with context variables, allowing a more flexible way to process semistructured data. In [19] the authors present a tool for verification of websites based on a subset of the Xcerpt language [6] and Abductive Logic Programming [18]. A detailed comparison between several approaches to verification can be found in [1].

5 Conclusion and Future Work

Our tool allows an easy development of rules with constraints to impose over XML content which can be used to automatically verify content in webpages submitted to open collaboration repositories such as Wikipedia. We believe that it can be further extended and used for example, as a browser plugin for constrained content presentation in the client side or to verify quality in terms of design and readability of a webpage.

Acknowledgements. Partially funded by LIACC through Programa de Financiamento Plurianual of the Fundação para a Ciência e Tecnologia (FCT).

References

1. Alalfi, M.H., Cordy, J.R., Dean, T.R.: Modelling methods for web application verification and testing: state of the art. Softw. Test. Verif. Reliab. 19(4), 265–296 (2009)
2. Alexandre, L., Coelho, J.: Filtering xml content for publication and presentation on the web. In: ICDIM, pp. 85–89. IEEE (2011)
3. Alexandre, L., Coelho, J.: Xcentric-based visual approach to web content verification. In: Simões, A. (ed.) Proceedings of 9th XML, Associated Technologies and Applications, pp. 71–82. Escola Superior de Estudos Industriais e de Gestão, Vila do Conde (2011)
4. Alpuente, M., Ballis, D., Falaschi, M.: A Rewriting-based Framework for Web Sites Verification. Electronic Notes in Theoretical Computer Science, pp. 41–61. Elsevier Science (2005)
5. Alpuente, M., Ballis, D., Falaschi, M.: Rule-based verification of web sites. STTT 8(6), 565–585 (2006)
6. Bry, F., Schaffert, S.: The XML Query Language Xcerpt: Design Principles, Examples, and Semantics. In: Chaudhri, A.B., Jeckle, M., Rahm, E., Unland, R. (eds.) NODe-WS 2002. LNCS, vol. 2593, pp. 295–310. Springer, Heidelberg (2003)
7. Bry, F., Schaffert, S.: Towards a Declarative Query and Transformation Language for XML and Semistructured Data: Simulation Unification. In: Stuckey, P.J. (ed.) ICLP 2002. LNCS, vol. 2401, pp. 255–270. Springer, Heidelberg (2002)
8. Coelho, J., Florido, M.: CLP(Flex): Constraint Logic Programming Applied to XML Processing. In: Meersman, R. (ed.) OTM 2004. LNCS, vol. 3291, pp. 1098–1112. Springer, Heidelberg (2004)
9. Coelho, J., Florido, M.: VeriFLog: A Constraint Logic Programming Approach to Verification of Website Content. In: Shen, H.T., Li, J., Li, M., Ni, J., Wang, W. (eds.) APWeb Workshops 2006. LNCS, vol. 3842, pp. 148–156. Springer, Heidelberg (2006)
10. Coelho, J., Florido, M.: Type-based static and dynamic website verification. In: The Second International Conference on Internet and Web Applications and Services. IEEE Computer Society (2007)
11. Coelho, J., Florido, M.: XCentric: logic programming for XML processing. In: ACM International Workshop on Web Information and Data Management (WIDM 2007), pp. 1–8 (2007)

12. Despeyroux, T.: Practical semantic analysis of web sites and documents. In: Feldman, S.I., Uretsky, M., Najork, M., Wills, C.E. (eds.) WWW, pp. 685–693. ACM (2004)
13. Foundation, W.: Wikipedia: Multilingual, web-based, free content encyclopedia project (2011) World Wide Web, http://www.wikipedia.org/
14. van Harmelen, F., van der Meer, J.: *WebMaster*: Knowledge-Based Verification of Web-Pages. In: Imam, I., Kodratoff, Y., El-Dessouki, A., Ali, M. (eds.) IEA/AIE 1999. LNCS (LNAI), vol. 1611, pp. 256–265. Springer, Heidelberg (1999)
15. Henzinger, M.R., Henzinger, T.A., Kopke, P.W.: Computing simulations on finite and infinite graphs. In: FOCS, pp. 453–462 (1995)
16. Kutsia, T.: Context sequence matching for XML. In: Proceedings of the 1st Int. Workshop on Automated Specification and Verification of Web Sites (2005)
17. Lloyd, J.W.: Foundations of Logic Programming, 2nd edn. Springer (1987)
18. Mancarella, P., Terreni, G., Sadri, F., Toni, F., Endriss, U.: The ciff proof procedure for abductive logic programming with constraints: Theory, implementation and experiments. TPLP 9(6), 691–750 (2009)
19. Mancarella, P., Terreni, G., Toni, F.: Web sites repairing through abduction. Electr. Notes Theor. Comput. Sci, vol. 235, pp. 137–152 (2009)
20. Mayol, E., Teniente, E.: A Survey of Current Methods for Integrity Constraint Maintenance and View Updating. In: Akoka, J., Bouzeghoub, M., Comyn-Wattiau, I., Métais, E. (eds.) ER 1999. LNCS, vol. 1728, pp. 62–73. Springer, Heidelberg (1999)
21. Microsoft: The C# Language (2011), http://msdn.microsoft.com/en-us/vcsharp/aa336809.aspx
22. Uwe Lesta, S.S.G.: A CSharp class library to connect. NET languages with SWI-Prolog (2011) World Wide Web, http://www.lesta.de/prolog/swiplcs/Generated/Help/introduction.htm
23. W3C: XML Schema (2010) World Wide Web, http://www.w3.org/XML/Schema/
24. Wielemaker, J.: SWI Prolog. WWW (2011), http://www.swi-prolog.org/
25. Wikipages (2011), http://www.wikipages.com

Histogram Correlation for Video Scene Change Detection

Nisreen I. Radwan[1], Nancy M. Salem[2], and Mohamed I. El Adawy[2]

[1] National Research Centre, Cairo, Egypt
eng_nesrin@hotmail.com
[2] Faculty of Engineering, Helwan University, Cairo, Egypt
nancy_salem@h-eng.helwa.edu.eg, mohamed_eladawy@cic-cairo.com

Abstract. In this paper a novel and simple scene change detection algorithm based on the correlation between the frames of the video is proposed. The first frame of the video is taken as a reference frame. The correlation between the histogram of the reference frame and the histogram of all video frames is computed. The plotting of the relationship between the computed correlation values and frame number illustrates the differentiation between scene and motion changes. When the correlation values are constant over a number of frames, so there is a motion scene where the background is not changed. While changing the correlation values over a number of frames indicate a gradual scene change. Changing of these values sharply indicates abrupt scene change. Experimental results show that this method is effective for motion, abrupt and gradual shot transition detection. It achieves an F-measure exceeding 0.89 for gradual shot transition compared with 0.84 when using a PCA based method.

Keywords: scene change detection, gradual transition, abrupt transition, image histogram, correlation.

1 Introduction

Digital video data type has been increasing rapidly in areas such as video conferencing [1], multimedia authoring systems [2], education and video on demand systems [3], [4]. Thus an effective method to find desired video information from a huge database using content is required. Video segmentation is the first step in video analysis for indexing, browsing and retrieval the video data. This segmentation process is generally called shot boundary detection or scene change detection. A shot is a sequence of frames generated during a continuous camera operation and represents a continuous action or a meaningful event. A scene is composed of a number of shots. The hierarchical description of video is shown in Fig.1 [5]. Scene transitions can be divided into two categories: abrupt transitions (cuts) and gradual transitions (fads, dissolves, and wipes).

Cut: It is a hard boundary coming from instantaneous change from one shot to another as shown in Fig. 2.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 765–773.
springerlink.com

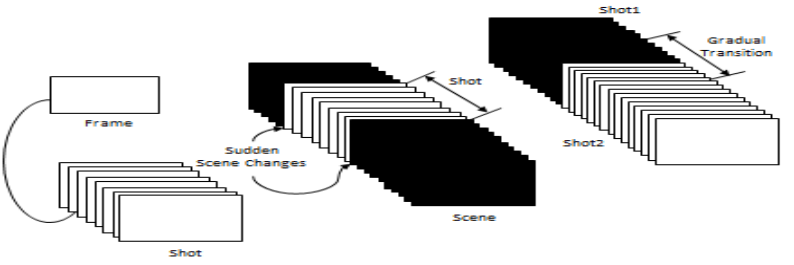


Fig. 1. A Hierarchical description of video sequence

Dissolve/Fading: In video production, proportion of two picture signals is added together so the two pictures appear to be merged on the screen. This process is used to move on from picture F to picture G. If the contribution of picture F changes from 100% to zero, and the contribution of picture G changes from zero to 100% then it is called a dissolve. When picture F is a solid colour, it is a fad-in and when picture G is a solid colour, it is a fad-out [5]. Figs. 3-5 show examples of dissolve and fading.

Wiping: This is a virtual moving boundary going across the screen clearing the old scene and displaying a new scene. This moving boundary can be a line or a set of lines [6]. Two kinds of wiping transition are shown in Figs. 6 and 7.

Fig. 2. Abrupt transition (cut)

Fig. 3. Dissolve (V_1)

Fig. 4. Fad-in (V_2)

Fig. 5. Fad-out (V_2)

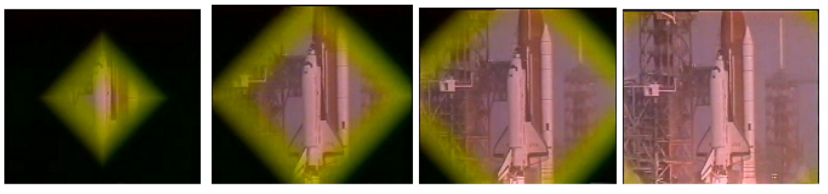


Fig. 6. Wipe1 (V_3)

Fig. 7. Wipe2 (V_3)

Many accurate algorithms have been developed for abrupt scene transition detection in recent years. However, the gradual transition detection is still a challenging problem for its need to drastic changes between two consecutive frames. Gradual transition also has a potential mixture with local object motion and global camera motion.

Zhang et al. [7] proposed a method based on the difference of intensity histogram of the frames and a dual threshold values were applied to detect abrupt and gradual shot transition. Meng et al. [8] proposed a method depend on the intensity variance of successive frames. A possibility that scene changes occur when a large depth of intensity variance valley is detected. Yeo and Liu [9] directly extracted DC images in the compressed domain and then identified shot transitions based on the difference of histograms between DC images. Zabih et al. [10] put a method based on calculating edge change fraction of every frame in temporal domain to detect cuts, fads, and dissolves. Qian and Liu [11] detected the fades based on the accumulating histogram difference (AHD) for both the compressed and uncompressed videos.

A supervised classification method and a framework for using different kinds of features extracted from the video for detecting various types of shot boundaries were introduced in [12]. A scene change detection algorithm based on neural network was introduced by Lee in [13], where the DC image was extracted and the variance was computed for every video frame. Feature vectors such as pixel-wise

difference, histogram difference and normalized correlation difference were used as input vector to the neural network. This algorithm showed promising results but also it had some limitations such as false detection when a big object was moving within the frame.

Some schemes based on audio and video content analysis were introduced in [14], [15]. By combining visual and audio boundary features, the scene change detection was enhanced. Gao [16] detected scene change by using the principle component analysis of video data. Where the difference between a one dimension PCA features for every two consecutive frame was computed. Zhi Li [17] proposed a scheme based on 3D wavelet transform. Where the 3D wavelet transform can effectively express the correlation of the several successive frames. Three features are computed over a window of frames to describe the correlation of the shot transitions. Table 1 summarizes the evaluation measure F_1 [12], [18] for 3 algorithms.

Table 1. F_1 evaluation measure for 3 schemes

Video scheme	*Gradual Transition*	*Abrupt Transition*
[12]	0.69	0.94
[18]	0.789	0.95
[16]	0.805	0.95

In these algorithms, both the correlation between two frames or over a window of frames is employed and the accuracy of the gradual transition detection is low. The number of frames in the window also, can't be indicated precisely. In this paper, a new scheme for shot transition detection is proposed. The scheme focuses on identifying the existence of a transition rather than its precise temporal kind. It depends on studying the histogram correlation between a reference frame and the rest of the frames in the video.

This paper is organized as follows. Section 2 describes the proposed scheme. Section 3 presents the experiment results. The conclusion and references are in Sections 4 and 5.

2 Histogram Correlation

Correlation is a single number indicates the degree of relationship between two variables. The video frames consist mainly of cuts, gradual transitions, and motions. For a cut, the dissimilarity of two neighbouring frames is strong, and the correlation of the frames is weak. For gradual transition, the two neighbouring frames are different in the pixel value, but similar in the edges and the texture, so the correlation in the spatial domain is very strong [17]. The changes in the intensity histogram of a motion scene which happen on the same background can be almost constant but for gradual transition and cuts it changes gradually or sharply. The histogram is a graphical representation showing the number of pixels belonging to each grey level in the frame.

Differentiation between object motion and scene transition can be obtained by taking a reference frame fr which is the first frame in the proposed algorithm. The histogram H_{fr} of the referenced frame is calculated as in [19]:

$$H_{fr}(r_k) = n_k \tag{1}$$

Where r_k is the kth intensity level and n_k is the number of pixels in the frame whose intensity level is r_k.

For N frames in the video file, we compute the histogram H_i where i=2,3,…, N frames of the video. The correlation between H_{fr} and H_i can be computed as follows:

$$corr(H_{fr}, H_i) = \frac{\sum_{j=1}^{n}(H_{fr}(j) - h_{fr})(H_i(j) - h_i)}{\sqrt{\sum_{j=1}^{n}(H_{fr}(j) - h_{fr})^2 \sum_{j=1}^{n}(H_i(j) - h_i)^2}} \tag{2}$$

Where n is number of gray scale levels r_k, h_{fr} and h_i are the means of H_{fr} and H_i. The algorithm is described in Fig. 8.

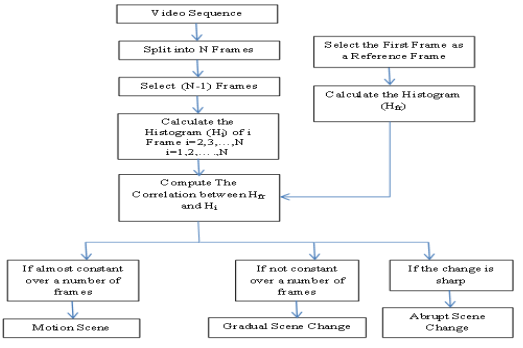


Fig. 8. The algorithm description

Most of previous histogram based algorithms depend on the comparison between each two consecutive frames in the video. Other algorithms make this comparison over a window of frames. The new in the proposed algorithm is the reference frame which makes a good differentiation between motion and scene transition. Three video sequences ("V_1", "V_2" and "V_3") are used to test the proposed algorithm. These videos are partials of videos of TREC video test repository [20]. Each one of these videos contains different kind of video transition. V_1 contains the dissolve transition, V_2 contains fad in and fad out, and V_3 contains the wiping transition as shown in Figs. 3-7. Fig. 9 shows the detection of the dissolve transition in Fig.3 which started at frame number 200 and finished at 238. Also, the cut can be detected as the straight line shown in the same figure. In Fig. 10, fad-in and fad-out can be detected. In Fig. 11, It is shown that wiping 1 can be detected while wiping 2 is not fully detected. This is because the different between the number of lines in each wiping.

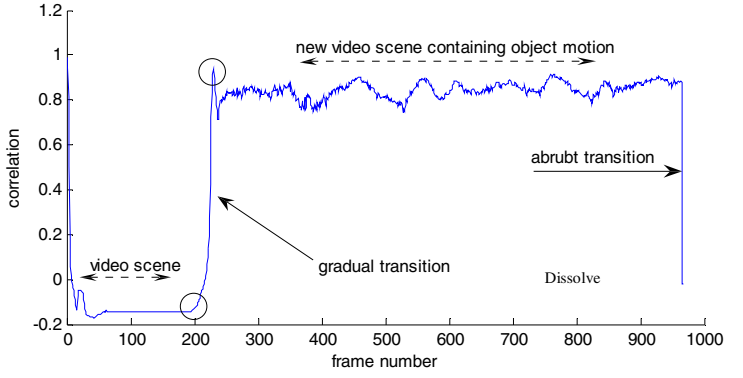


Fig. 9. Detection of dissolve transition in V_1

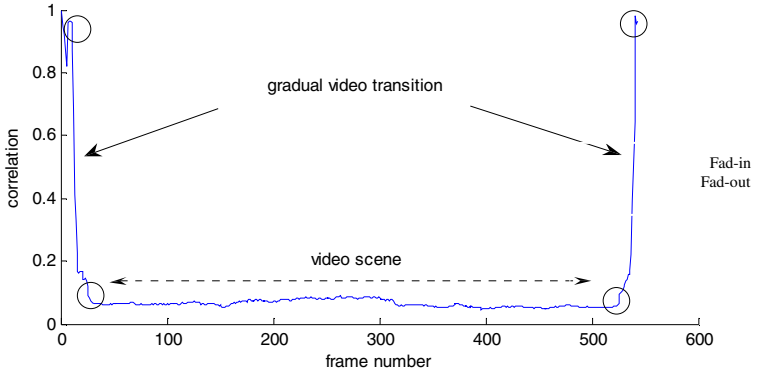


Fig. 10. The detection of fad-in and fad-out

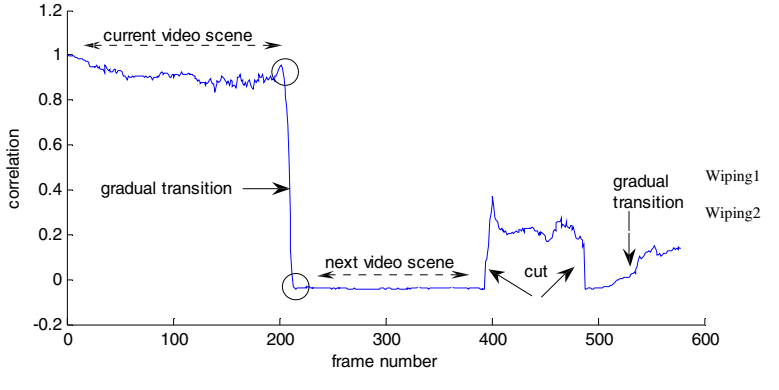


Fig. 11. The detection of wiping

The algorithm steps:

1. Split the video sequence into N frames.
2. Take the first frame as a reference frame. Calculate its histogram H_{fr}.
3. For the remainder frames, calculate the histogram H_i.
4. Compute the correlation between H_i and H_{fr}.
5. Plot the relationship between the correlation and frame number to indicate the scene and motion changes.

3 Experimental Results

In our experiments, the video test sequence "NAD30" is used for performance evaluation of the proposed algorithm. The number of frames used was 10000 frames. In most scene change detection algorithms, the recall and precision are the two commonly used measures. These measures are given by:

$$P_{re} = \frac{N_c}{N_c + N_m} \tag{3}$$

$$P_{pre} = \frac{N_c}{N_c + N_f} \tag{4}$$

Where P_{re} is the recall and P_{pre} is the precision. The number of the missed detections is N_m , the number of false alarms is N_f and the number of correctly detections is N_c .

F_1 is a commonly measure that combines both recall and precision [12], [18] which is given in Eq. (5) as:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall} \tag{5}$$

Table 2 describes F_1 for the proposed algorithm which achieves high result for gradual transition detection while, a reasonable result for abrupt scene transition. In this table, our results were compared with the method proposed in [16] when using the same video sequence "NAD30".

Table 2. F_1 evaluation measure for the proposed scheme

Video Scheme	*Gradual Transition*	*Abrupt Transition*
proposed	0.892	0.877
Gao [16]	0.844	0.842

4 Conclusions

A simple correlation based scene change detection algorithm was presented. The first frame of the video sequence has been taken as a referenced frame. The histogram of

the referenced frame and all the remained frames was calculated. The correlation between the histogram of the referenced frame and the histogram of remaining video frames was calculated to indicate the scene changes. Experiments were carried out on the TREC video test repository. Results show that the F-measure was 0.892 and 0.877 for gradual and abrupt transitions respectively using our proposed algorithm.

References

1. Guoping, Y., Lijuan, H.: Design and Implementation of the SIP Video Conferencing System in Public Security. In: IEEE International Conference on Multimedia Technology, ICMT (2010)
2. Rizzo, G., Meirone, B.: Distributed Semantic Video Tagging for Peer-to-Peer Authoring System. In: IEEE Workshop on Database and Expert Systems Applications, DEXA (September 2010)
3. Ma, L., Shen, H., Zhang, Q.: The Key Technologies for a Large-Scale Real-Time Interactive Video Distribution System. In: IEEE International Con-ference on Advanced Computer Control, ICACC (2010)
4. Semertzidis, T., Dimitropoulos, K., Koutsia, A., Grammalidis, N.: Video Sensor Network for Real-time Traffic Monitoring and Surveillance. In: IEEE International Conference on Intelligent Transport Systems, IET (2010)
5. Fernando, W.A.C., Canagarajah, C.N., Bull, D.R.: A Unified Approach to Scene Change Detection in Uncompressed and Compressed Video. IEEE Transaction on Consumer Electronics 46(3) (August 2000)
6. Adhikari, P., Gargote, N., Digge, J., Hogade, B.G.: Abrupt Scene Change Detection. World Academy of Science, Engineering and Technology (2008)
7. Zhang, H., Kankanhalli, A., Smoliar, S.: Automatic Partitioning of Full-motion Video. In: ACM/Springer Multimedia Systems, pp. 10–28 (July 1993)
8. Meng, J., Juan, Y., Chang, S.: Scene Change Detection in a MPEG Compressed Video Sequence. In: Proc. SPIE, vol. 2419, pp. 14–25 (1995)
9. Yeo, B., Liu, B.: Rapid Scene Analysis on Compressed Video. IEEE Transactions on Circuits and Systems for Video Technology 5, 533–544 (1995)
10. Zabih, R., Miller, J., Mai, K.: A Feature-based Algorithm for Detecting and Classifying Scene Breaks. In: Proc. ACM Multimedia, San Francisco, pp. 189–200 (November 1995)
11. Qian, X., Liu, G.: Effective Fades and Flashlight Detection Based on Accumulating Histogram Difference. IEEE Transactions on Circuits and Systems for Video Technology 16(10), 1245–1258 (2006)
12. Qi, Y., Hauptmann, A., Liu, T.: Supervised Classification for Video Shot Segmentation. In: Proc. IEEE Conf. on Multimedia Expo. (ICME), vol. 2, pp. 689–692 (2003)
13. Lee, M.-H., Yoo, H.-W., Jang, D.-S.: Video Scene Change Detection using Neural Network: Improved ART2. Expert Systems with Applications 31, 13–25 (2006)
14. Zhu, Y., Zhou, D.: Scene Change Detection Based on Audio and Video Content Analysis. In: Proceedings of the Fifth International Conference on Computational Intelligence and Multimedia Applications (ICCIMA 2003), IEEE (2003)
15. Kyperountas, M., Kotropoulos, C., Pitas, I.: Enhanced Eigen-Audio frames for Audiovisual Scene Change Detection. IEEE Transactions on Multimedia 9(4) (June 2007)

16. Gao, L., Jiang, J., Liang, J., Wang, S., Yang, S., Qin, Y.: PCA-based Approach for Video Scene Change Detection on Compressed Video. IEEE Electronic Letters 42(24) (November 2006)
17. Li, Z., Liu, G.: A Novel Scene Change Detection Algorithm based on the 3D Wavelet Transform. In: IEEE International Conference on Image Processing, ICIP, pp. 1536–1539 (2008)
18. TREC Video Retrieval Evaluation (2005), http://www.nlpir.nist.gov/projects/trecvid/
19. Gonzalez, R.C., Woods, R.E., Eddins, S.L.: Digital Image Processing Using Matlab. Pearson Prentice Hall, New Jersey (2004)
20. The open Video Project, http://www.open-video.org

Microposts' Ontology Construction

Beenu Yadav[1], Harsh Verma[2], Sonika Gill[3], and Prachi Bansal[4]

[1] M. Tech (CS) Student
Radha Govind Group of Institutions, Meerut
beenu_yadav@rediffmail.com
[2] M. Tech (CS) Student
Radha Govind Group of Institutions, Meerut
er.harsh86@gmail.com
[3] M. Tech (CS) Student
Radha Govind Group of Institutions, Meerut
sonika.gill108@gmail.com
[4] M. Tech (CS) Student
Radha Govind Group of Institutions, Meerut
enggprachi@gmail.com

Abstract. The social networking website Facebook offers to its users a feature called "status updates" (or just "status"), which allows users to create Microposts directed to all their contacts, or a subset thereof. Readers can respond to Microposts, or in addition to that also click a "Like" button to show their appreciation for a certain Micropost. Adding semantic meaning in the sense of unambiguous intended ideas to such Microposts. We can make a start towards semantic web by adding semantic annotation to web resources. Ontology are used to specify meaning of annotations. Ontology provide a vocabulary for representing and communicating knowledge about some topic and a set of semantic relationships that hold among the terms in that vocabulary. For increasing the efficiency of ontology based application there is a need to develop a mechanism that reduces the manual work in developing ontology. In this paper, we proposed Microposts' ontology construction.

Keywords: Microposts, Lexicon, Sysnset, Universal Decimal Classification (UDC), Statistically Indexed Table, Ontology, Concept Extraction, Syntatic Parsing.

1 Introduction

Social media offers a great medium for people to share their opinions and thoughts, which in turn provides a wealth of useful information to companies and their rivals, other consumers and analysts. While finding out what a single person likes and dislikes is not particularly useful on its own, the associations and conclusions that can be drawn from finding and clustering groups of people with similar interests is a veritable goldmine, going from the direct: "this group of people likes Nike products", to the indirect: "People who like skydiving tend to be risk-takers", to the associative: "People who buy Nike products also tend to buy Apple products". However, the difficulty lies in

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 775–788.
springerlink.com

accurately extracting the relevant information from the text: this is problematic even from well written sources such as online newspapers, articles and reports, but more difficult still from social media such as blogs, twitter, facebook and so on, where people use slang, do not write in full sentences or correct English, and make assumptions about the world knowledge of the reader, for example about popular culture such as books, films, news items and so on. Furthermore, it can be difficult even for a human to understand the finer concepts of the use of irony and sarcasm which is particularly present in social media, let alone for a machine.

While there are a number of sentiment analysis tools available which summarise positive, negative and neutral tweets about a given keyword or topic, these tools generally produce poor results, and operate in a fairly simplistic way, using only the presence of certain positive and negative adjectives as indicators, or simple learning techniques which do not work well on short Microposts.[4]

An ontology defines a common vocabulary for researchers who need to share information in a domain. It includes machine-interpretable definitions of basic entities in the domain and relations among them.[9] We develop ontology due to following reasons:

- To share common understanding of the structure of information among people or software agents
- To enable reuse of domain knowledge
- To separate domain knowledge from the operational knowledge
- To analyze domain knowledge.

1.1 Defining Ontology

Ontology is an explicit formal specification of the terms in the domain and relations among them.

Ontology is a formal explicit description of [7]:

- ***Semantic Relations*** among concepts
- ***Concepts*** in a domain of consideration (called classes or concepts)
- ***Properties*** of each concept called concept description.
- ***Restrictions*** on properties also called facets.

A *concept* is an abstract, universal idea, notion or entity that serves to designate a category or class of entities, events or relations. It is a mental picture of a group of things that have common characteristics. *Classes* delineate concepts in the domain so they are the focus of most ontology. *Semantic relations* depict the collaboration of two concepts. *Properties* describe various features and attributes of the concept. Properties can have different *restrictions* such as value type, allowed values, number of values and other features of the values the property can take.

In practical terms, *Ontology construction* includes:

- Defining classes in the ontology,
- Relating the classes with a semantic relation,
- Arranging the classes in a taxonomic (subclass–superclass) hierarchy,
- Defining properties and describing allowed values for them,
- Filling in the values for properties for instances.

We can then create a knowledge base by defining individual instances of these classes filling in specific attribute value information and additional property restrictions.

"An ontology together with a set of individual instances of classes constitutes a knowledge base" [7].

1.2 Ontology Design

The ontology includes concepts and semantic relations with other concepts of the same domain. The concepts are described as a class, which includes their properties and restrictions on the values of the properties. The subclass inherits all the properties of the superclass but does not inherit the relationships with other classes.

1.2.1 Ontology Schema

Ontology is a specification of semantically related concept nodes. Ontology Schema can be represented by the structure of a concept node.

Concept ID: It is a unique identification of the Concept. The Concept Id is represented by any universally acceptable identification scheme. For the ease of understanding presently we are using a unique integer for concept identification such as C#110 is the Id for concept TCP/IP.

Concept ID
Concept Name
Generic Properties
Class Specific Properties
Semantic Relations between Concepts
Restrictions

Concept ID – C# 110
TCP/IP
Is the most popular open-system protocol suite for communication.
Is Robust.
Connects: NETWORKS, Detects: ERRORS, Composed_of: LAYERS
Null

Concept Node with Example

Concept Name: It signifies name of class corresponding to the Concept Id. Concept is a general idea formed in the mind. It is an idea about a group of things. A concept involves thinking about what it is that makes those things belong to that one group. Each word in the input text belongs to a group that identifies the concept.

Generic Properties: A set of attributes, settings and/or parameters used to define or describe an object. If a class1 has IS_A relationship with class2 it implies that it is a subclass of class2. Class1 will inherit all the properties of class2.

Class Specific Properties: Each class has its own properties defining its attributes.

Semantic Relations between concepts: This defines the relationship of a concept with others concepts. A concept may not be related with every other concept in Ontology.

Restrictions: The types of restrictions which can be imposed in an ontology can be categorized as:

- Language Constructs: these restrictions exist on property only and the methods to represent restrictions on property are given in Web Ontology Language and are named as *Property Restrictions* and *Restricted Cardinality* [11].
- Restriction on Concepts: defined by quantifiers such as double, one-fifth etc. For example, if somewhere we talk about one-third of population then 'POPULATION' is a concept with restriction one third. It is because we are considering only one-third population instead of entire population.
- Restriction on Semantic Relation: defined by conditional sentences. For example, if the sentence is,
  If Aditya will talk Mary, then he will meet with Alice.
  In this sentence, the relationship 'will_meet' between the concepts ADITYA and ALICE exists with the constraint 'If Aditya will talk Mary'.

2 Defining Vibhakti Parser

The parser verifies the grammatical correctness of the input text and identifies the 'Vibhaktis' or 'Case Roles' in the input text. So we call it *"Vibhakti Parser"*. The Vibhakti Parser performs two functions.

- Parsing the text
- Identifying the Vibhaktis/Case Roles

2.1 Parsing the Text

To parse the text, parser uses language grammar rules [1, 11], which are defined as production rules. This parsing examines the syntax of the text and results that text is syntactically correct or incorrect.

Parser is a collection of rules for representation of sentences in the form of production rules. The Production rules can be written as,

<*simple sentence*> = <*subject*> < *verb*> <*complement*>

The Parser has production rules for all types of sentences such as Simple sentences, Compound sentences etc.

2.2 Identifying the Vibhaktis/Case Roles

Within a sentence different nouns are connected with verb through case relationship. To identify these case relations in each language vibhaktis are used. The Paninian Grammar Framework concerns the Sanskrit language [13, 10]. However, it prescribes a generic and language independent decomposition of any sentence into eight different information carrying vibhaktis. These vibhaktis or case roles are as follows:

1. ***Kartaa/ Nominative*** - Doer of an activity or the subject.
2. ***Karma/Accusative -*** Entity that is being acted upon or the object.
3. ***Karan/Instrumental -*** Entity that is being employed to complete an act.
4. ***Sampradan/Dative*** - The chief motivation behind the action of the beneficiary subject.
5. ***Apadan/Ablative -*** Entity in Karma is separated as a consequence of the action.
6. ***Sambandh/Genitive*** - The possessor of something in the sentence.
7. ***Adhikaran/Locative -*** Place, time related to the entity at the time of action.
8. ***Sambodhan/Vocative*** - Calling upon someone – hey etc.

For example, consider the sentence,

English: The student presented the seminar of his project with projector in seminar hall.

Hindi: Student ne Apne Project ka Seminar Kaksha mein Projector se seminar ko present kiya

In this sentence,

(i) Student – *Kartaa*
(ii) Seminar – *Karma*
(iii) Projector – *Karan*
(iv) His Project – *Sambandh*
(v) Seminar Hall – *Adhikaran*

2.3 Syntactic Parsing

Syntactic parsing examines the sentence syntactically and results valid sentence, if sentence is syntactically correct else results invalid sentence. The language grammar rules, which are defined in the form of production rules, are used to parse the text [1, 5]. For representation of sentences, production rules are described in the parser. It includes representation for all types of sentences. Input sentences are parsed by defined sentence structure rules and when it sets to any one of the rules then that sentence is proved to be syntactically correct.

Example:

S1: I called him but he gave me no answer.

→ <Simple Sentence> <Conjunction> <Simple Sentence>
→ <I> <called him> <Conjunction> <he> <gave me no answer>
→ <subject1> <predicate1> <Conjunction> <subject2> <predicate2>

<subject1> → <nominative personal pronoun>
<predicate1> → <V> <complement>
→ <Vpast> <object>

<subject2> → <nominative personal pronoun>
<predicate2> → <V> <complement>
→ <Vpast> <indirect object> <object>

2.4 Vibhakti Parsing

The Vibhakti Parser parses the syntactically correct sentence to identify the vibhaktis, states, verbs and others elements. The rule base is made for determination of each of them. After remodeling we apply the following rules and identify Vibhaktis, States, etc.

2.4.1 Rule Base

For identification of Vibhaktis/Case roles

1. Subject of the sentence is identified as Kartaa Vibhakti.
2. If the subject has pronoun then Parser replace it with the corresponding noun, it is identified as Kartaa Vibhakti.
3. Rest of the Vibhkatis are identified from complement of the sentence.
    a. If complement has an object(direct/indirect) then it is Karam Vibhakti.
    b. In case of pronoun object before determining Vibhakti, Parser substitutes it with its respective noun.
4. The vibhaktis are identified by preposition in the prepositional phrase.
5. In prepositional phrase if
    a. Preposition is *" Main verb+ to + NP "* → Karam Vibhakti
    b. Preposition is *"by, with, from"* → Karan Vibhakti
    c. Preposition is *"for, to + Vinf"* → Sampradaan Vibhakti
    d. Preposition is *"from*, by*"* → Apadaan Vibhakti
    e. Preposition is *"of, to*"* → Sambandh Vibhakti
    f. Preposition is *"at, in, on, above"* → Adhikaran Vibhakti

from* => 'from' when used with some special verbs that indicate separations such as fell, break or some phrases as fell down etc. then it is categorized as Apadaan Vibhakti else it is Karan Vibhakti.

by* => 'by' when used with some special verbs that indicate separations such as fell or some phrases as letting off etc. then it is categorized as Apadaan Vibhakti else it is Karan Vibhakti.

to* => 'to' when used in the form other than as explained in 'a' and 'c' then it is Sambandh Vibhakti.

We have categorized some prepositions for identifying Vibhaktis/Case roles. In a similar manner this categorization of prepositions can be enhanced by working on more prepositions such as compound prepositions, phrase prepositions.

For identification of Verbs

1. Verbs or verb phrases in the sentence represent actions.

For identification of States

1. Some sentences represent state rather than actions; the state is identified as property of the subject.

For identification of Other Elements

1. The conditional sentences impose restrictions on either the verbs or the property. The 'if' clause or 'when' clause of such sentences is added to all the relations.
2. The quantifiers are added as restrictions to the noun/noun phrase that will be further identified as concepts in the construction of ontology.

2.5 Formation of Vibhakti Table

The Vibhakti Parser generates the Vibhakti Table of the input document on applying vibhakti parsing rules on syntactically correct simple sentences. Vibhakti Table has columns for Verb of the sentence, one for property of Kartaa in the sentence, seven for Vibhaktis/case roles of sentence. Using the above defined rules, Vibhakti Parser frames a Vibhakti Table for given text/document.

2.5.1 Steps for Framing Vibhakti Table

1. Each sentence is processed for syntactic correctness by using Production rules defined above in Syntactic Parsing section.
    a. If the parsed sentence (after remodeling, if any) is valid in grammatical sense then it undergoes Vibhakti Parsing.
    b. Else Syntactic Parsing is interrupted and the subsequent sentence is treated as the next input for parsing.
2. Each syntactically valid simple sentence is scanned for identifying noun phrases, verbs or prepositional phrases. As the Parser encounters any one of these then using Vibhakti Parsing rules, Vibhaktis/case roles, verbs and properties are determined.

The determined vibhaktis, verbs and properties are simultaneously fed into the respective cell of Vibhakti Table.

Example:

- *The lecture was focused on the problem of unemployment.*

Vibhakti/Case Role Table

S. No.	*Verb*	*Kartaa*	*Karam*	*Karan*	*Sampradan*	*Apadan*	*Sambandh*	*Adhikaran*	*Property*
1	Was focused	The lecture					of unemployment	on the problem	

3 Concept Extractor

The concept extractor is a module designed for the determination of concepts of the ontology. *The nouns and the noun phrases are the keys which form concepts in the ontology* [8, 2, 12]. For this purpose we scale some existing linguistic resources according to our requirement and design new components using some existing resources.

3.1 Lexicon

A Lexicon is a repository of words and knowledge about those words. A lexicon is a list of words together with additional word-specific information. It is a list of corresponding terminology in different languages, usually locale, industry or project specific [3].

Lexicon used for microposts ontology builder, incorporates-

1. Collection of Words
2. Unique Id(s) respective to each word: It is a Universal Decimal Classification (UDC) that uniquely identifies the concepts. The UDC(s) are determined from the SynSet table.
3. The category to which the word belongs based on classification of concepts is attached. The classification of concepts is given in the forthcoming section.

The word extracted from text/document for the identification of concept may or may not be matched with any word from the collection of words in Lexicon. When word does not match with any entry of Lexicon directly then morphology [6] is used.

For Example, words like Networks, Leaves etc., are not found in Lexicon. In these words morphemes are –

1. Network, -s
2. Leaf, -ves

To identify UDC(s) for these words, these words are analyzed as sequence of morphemes so that one of the word forms gets matched in Lexicon.

3.2 SynSet Table

The SynSet Table is a table developed for the identifications of words possessing the same meaning. It is the collection of synonymous words with the attribute set. The unique identification number is given to the set of words that have the identical meaning and such set identify the unique concept.

To each unique concept we give *UDC (Universal Decimal Classification)* identification as its unique identification number. The UDC is the world's foremost multilingual classification scheme for all fields of knowledge. An advantage of this system is that it is infinitely extensible, and when new concepts are introduced, they need not disturb the allocation of numbers to the existing concepts [13].

In every language there are some words that express multiple meanings when used in different contexts. The exact meaning of such word is determined from the context of sentence in which the word is used. For this purpose we attach an attribute set with such words in the SynSet Table. In case when a word with different meaning in different contexts is encountered then the attribute set is exploited for the identification of exact word.

Each row in the SynSet table consists of three columns.

a) The first column of every row has UDC.
b) The second column has synonymous words having the same concept.
c) The third column has Attribute Set. The motivation for this is to provide a framework for finding semantically sensible concept of a multi-contextual word provided by the Lexicon.

For Example,

UDC	Synonym Set	Attribute Set
5/6:523.31.12	Space, Area, Volume, Region	one, two, or three dimensional; bounded, occupied by objects
5/6:528.93	Space, Outer Atmosphere	Related to solar system, beyond the earth's atmosphere, boundless

3.3 Statistically Indexed Concept Table

Extracting concepts requires a technique that can retrieve the appropriate concepts from documents of any subject domain. Statistical indexing technology is accurate enough to compute extraction of concepts [2].

The Vibhakti Parser extracts the units, such as noun phrases; they can be used to depict concepts by computing their frequency across the document. The indexing can be accomplished by computing the statistical frequency of extracted noun phrases within each document in a collection. The Statistically Indexed Concept Table is constructed by entering each noun phrase with its UDC. The UDC is determined from Lexicon and SynSet table. The noun/noun phrases, their UDC identification and their count altogether shape the Statistically Indexed Concept Table.

Example: The Statistically Indexed Concept table

Row No.	Nouns/Noun Phrases	Frequency	UDC
1	TCP/IP, TCP and IP	7	681.324.003
2	Local Area Network, LAN, LAN operations	3	681.324.001
3	Computer Networks	5	681.324

The frequency index of each noun/noun phrase changes while the document is read. The frequency index of the table corresponding to each concept determines the validated concepts of the ontology.

3.4 Concept Extraction Method

This section outlines the methodology for figuring out the concepts for an ontology using above illustrated components and resources. Lexicon and SynSet Table are used to develop the Statistically Indexed Concept table, which is used to determine the concepts for the ontology. The step wise procedure is given as:

1. The word/phrase is extracted from the sentence to determine its concept.
2. This extracted word/phrase is mapped to the Lexicon. The Lexicon consists of UDC(s) relative to each word. These Unique Id(s) is used to find the concept(s) from SynSet table.
3. There may be more than one Unique Id corresponding to each word, which indicates that the word is used in different senses or contexts. The context of the extracted word is resolved using Attribute Set which is defined in SynSet Table.
4. The Unique Id found by the concept extractor is searched into the Statistically Indexed Concept Table. If it is found then the frequency corresponding to that Unique Id is increased by one and the extracted noun/noun phrase is appended to the Noun/Noun Phrase column.
5. For each extracted word/phrase

a) If the extracted word/phrase has one UDC in the Lexicon then this identification is fed into Statistically Indexed Concept Table.

b) Otherwise the complete sentence is read and the SynSet table is referred to determine its unique concept. With the help of Attribute Set and the sentence, the unique concept of the word/phrase is determined. Corresponding to the unique concept the UDC is identified and fed into the Statistically Indexed Concept Table.

c) Unique Id and the extracted noun/noun phrase are made as a new entry into the table with the frequency 1.

4 Microposts' Ontology Builder

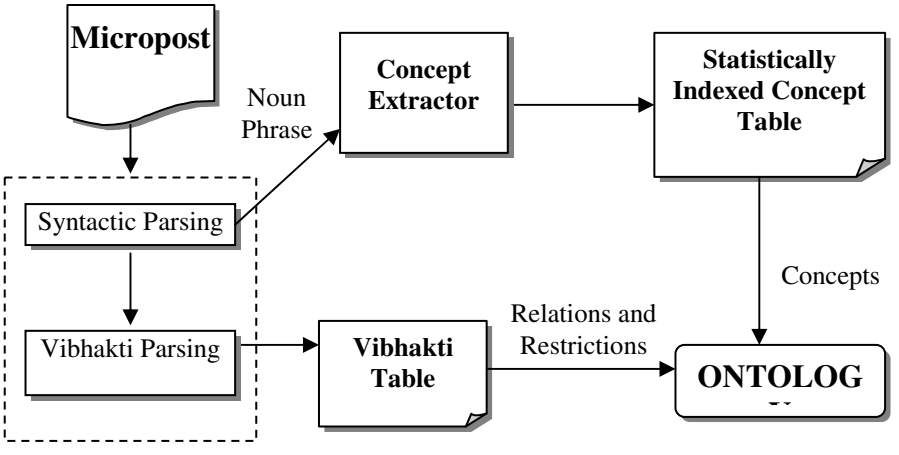
The Microposts' ontology builder is an endeavor to reduce the manual effort in the construction of ontology. This saves the time and thus efficiency of the work will be increased. We have explained the Vibhakti Parser which is a pillar of the Microposts' auto ontology builder. The second pillar of Microposts' auto Ontology Builder is

Concept Extractor. *Vibhakti Parser* with the *Concept Extractor* is integrated to develop ontology of any document. The forthcoming sections explain methodology for Microposts' ontology construction.

4.1 Architecture of Microposts' Ontology Builder

The development of Microposts' Ontology Builder is an approach to the automatic construction of ontology from the existing information resources.

The input document is passed to the Vibhakti Parser for the syntactic checking of the sentences and the noun/noun phrases identified during parsing are fetched by Concept Extractor to construct Statistically Indexed Concept Table. The Vibhakti table is constructed using the rule base of Vibhakti Parser. The concepts for the ontology under construction are determined from the Statistically Indexed Concept Table. These concepts and the Vibhakti Table, concurrently gives the structure to the ontology.



Architecture of Microposts' Ontology Builder

4.2 Functioning of Microposts' Ontology Builder

4.2.1 Algorithm

Step 1: **Parsing and Remodeling of the Sentence**

The input text/document is parsed for checking the grammatical correctness of the sentences and simultaneously the non simple sentences encountered are converted into simple sentences. The result of syntactic parsing and remodeling is syntactic tagged sentence and it is directly used for vibhakti parsing and for concept identification.

Step 2: **Vibhakti Parsing and Concept Identification**

The syntactically parsed sentence is used by Vibhakti Parser and Concept Extractor. On every tagged part of the sentence,

- the rules of vibhakti parsing are applied to identify the vibhaktis and
- simultaneously the noun/noun phrase are passed to concept extractor for the identification of concepts.

Step 3: **Construction of Statistically Indexed Concept Table**
The noun/noun phrase of the parsed sentence is used to identify concepts. The concept extractor uses Lexicon and SynSet Table to generate Statistically Indexed Concept Table, which contain the Unique Id and Frequency of occurrence corresponding to each concept.

Step 4: **Construction of Vibhakti Table**
The noun/noun phrase in the corresponding vibhakti column forms a concept and has an unique record in Statistically Indexed Concept Table. The noun/noun phrase and their respective Row No. retrieved from the Statistically Indexed Concept Table are fed into the vibhakti table.

The verbs of the sentence define the action, which is inserted into verb column of the Vibhakti Table.

The states are represented by properties, which is inserted into property column of the table.

The conditional sentences from the text impose the constraint on the action so it is written into the verb column of the row.

The quantifiers, multipliers etc. impose the restrictions on the nouns, which are fed into the Vibhakti column corresponding to that concept.

The Vibhakti Table identifies the vibhaktis, verbs, restrictions and properties such as dates, digits, units, formulae etc. Hence, Concept Extractor determines concepts and Vibhakti Parser parses each sentence of the text to construct the Vibhakti Table, which is ideally developed for the microposts' construction of ontology.

Step 5: **Approving the Concepts**
Since there are many concepts in the text of which ontology is to be made, out of all those some selected concepts will form the ontology, such selected concepts are approve concepts. Concepts are approved based on following procedure.

To approve concepts we refer to the statistically indexed concept table. This table has concepts with their UDC and the frequency of occurrence of concept in the input document. The concepts with the frequency index greater than the threshold value are approved concepts of the ontology to be built. The threshold value is determined beforehand. This value is application dependent and based on the criterion specified by the user.

Step 6: **Microposts' Ontology Formation**
Ontology is a specification of semantically related concept nodes. Ontology Schema can be represented by the structure of a concept node. For each approved concept identified from Kartaa Vibhakti we write a concept structure. A concept node structure includes:

- Concept ID
- Concept Name

- Properties
- Semantic Relations
- Restrictions

The Kartaa column of each row of the Vibhakti table is scanned subsequently to check that the noun/noun phrase is an approved concept. The elements that give structure to concept node relative to the approved concept are identified from the row of Vibhakti table. Otherwise the row of the Vibhakti table under consideration is not scanned further and the next row is scanned.

Concept ID and Concept Name

The concept Id is unique UDC identification taken from Statistically Indexed Concept Table. The name of the concept structure is the concept name, which is the highly significant noun/noun phrase retrieved from the respective column of the Statistically Indexed Concept Table.

Properties and Semantic Relations

The properties are written in sentential form. The properties that have a subset-superset type structure such as Is_a, Kind_of, Type_of followed by noun only or an adjective and a noun only then it forms a subset relationship which is included in semantic relations of the concept node.

The semantic relations in the ontology are identified from the vibhakti table with the help of verbs and the prepositions. For the determination of relationship here we state the semantics for writing the relations between concepts.

i. The relationship is determined from the main verb and the preposition.
ii. If the 'Sampradan' column of the row under consideration has verb then the relationship is identified by the verb in this column instead of combination of main verb and the preposition.
iii. If the row has an entry in 'Karam' column along with entries in other columns except 'Sampradan' then the relationship is identified by the combination of main verb, entry in 'Karam' column and the preposition.
iv. Relation between concepts that form Self loop is ignored unless the concepts have the restrictions/facets attached to them.

There may be instances when the approved concept is related to rejected concept but relationship between such concepts is included in the concept structure of the ontology built automatically.

Restrictions

1. Restriction on Semantic Relationship: The restriction on semantic relation is written with relationship in the concept structure.

2. Restriction on Concept: Constraints on concepts are portrayed in two forms.

- Based on the approved concept which has its concept structure.
  - If all the relations and properties are with same restricted concept then we write restriction with the concept name.
  - Else we categorize the relations and properties based on the restriction on the concept. The restriction is written with the categories.

- Based on the unapproved concept to which the concept node is related with a semantic relation.
  - The restriction is written with the unapproved concept.

Similarly, the entire table is scanned and the ontology of the text is constructed.

5 Conclusion

This paper proposed a technique to extract concepts from plain text to build ontologies. The extraction is based on existing linguistic resources like lexicon and synset. A Universal Decimal Classification is associated with each concept to classify the concepts. The Syntactic Parsing is to be done using Vibhakti Parser to preprocess the text and convert the compound and complex sentences into simpler sentences. The noun/noun phrases are extracted from the preprocessed text which are input to the concept extractor which extracts the potential nouns as the concepts. It uses Statistically indexed table is generated with the validation of the concept in text. Those concepts are extracted which are occurring most frequently in the text. This technique helps to extract the concepts from the Microposts'.

References

[1] Basic English Sentence Structures, http://www.scientificpsychic.com/grammar/enggram3.html
[2] Schatz, B.R.: The Interspace: Concept Navigation Across Distributed Communities. IEEE Computer (2002), http://www.canis.uiuc.edu/archive/papers/interspace.computer.pdf
[3] Lexicon, http://en.wikipedia.org/wiki/Lexicon
[4] Hartl, M.: Ruby on Rails Tutorial, http://ruby.railstutorial.org/chapters/user-microposts
[5] Modern English Grammar, http://papyr.com/hypertextbooks/grammar/
[6] Morphology (Linguistics), http://en.wikipedia.org/wiki/Morphology_%28linguistics%29
[7] Noy, N.F., McGuinness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology, http://www-ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf
[8] Bennett, N.A., He, Q., Chang, C., Schatz, B.R.: Concept Extraction in the Interspace Prototype. Technical Report, Digital Library Initiative Project, University of Illinois at Urbana-Champaign (1999), http://www.canis.uiuc.edu/archive/techreports/UIUCDCS-R-99-2095.pdf
[9] Ontology Working Group, http://mged.sourceforge.net/ontologies/index.php
[10] Sanskrit Grammar: Noun Cases, http://www.everything2.com/index.pl?node_id=1017898
[11] OWL Web Ontology Language Reference, W3C Recommendation (2004), http://www.w3.org/TR/2004/REC-owl-ref-20040210
[12] Vintar, S., Volk, P.B.M.: Semantic Relations in Concept-Based Cross-Language Medical Information Retrieval (2003), http://www.dcs.shef.ac.uk/~fabio/ATEM03/vintar-ecml03-atem.pdf
[13] UDC Consortium, http://www.udcc.org/

A Comparative Study of Clustering Methods for Relevant Gene Selection in Microarray Data

Manju Sardana and R.K. Agrawal

School of Computer and System Sciences, Jawaharlal Nehru University,
New Delhi, India 110067
manjusardana12@yahoo.co.in, rkajnu@gmail.com

Abstract. Classification of microarray cancer data has drawn the attention of research community for better clinical diagnosis in last few years. Microarray datasets are characterized by high dimension and small sample size. Hence, the conventional wrapper methods for relevant gene selection cannot be applied directly on such datasets due to large computation time. In this paper, a two stage approach is proposed to determine a subset containing relevant and non redundant genes for better classification of microarray data. In first stage, genes were partitioned into distinct clusters to identify redundant genes. To determine the better choice of clustering algorithm to group redundant genes, four different clustering methods were investigated. Experiments on four well known cancer microarray datasets depicted that hierarchical agglomerative with complete link approach performed the best in terms of average classification accuracy for three datasets. Comparison with other state-of-art methods have shown that the proposed approach which involves gene clustering is effective in reducing redundancy among selected genes to provide better classification.

Keywords: Clustering, Microarray, Gene Selection, Representative Entropy.

1 Introduction

DNA microarray technology has empowered the biologist to measure expression levels of thousands of genes simultaneously. One of the most common applications of microarray data is classification of samples into healthy versus diseased or normal versus abnormal by comparing gene expression levels. Microarray data which is characterized by high dimension and small sample size suffers from curse of dimensionality [1]. In order to design classifiers with good generalization capabilities along with minimum complexity burden associated with learning algorithm, there is need to reduce the dimension of microarray dataset. Feature extraction and feature selection are two common methods to reduce the dimension of data. Feature Extraction method transforms a given set of measurements to a new set of features. Suitable choice of transform can provide high information packing properties in new obtained features in comparison to original features. Alternatively, feature selection methods attempt to reduce the dimensionality by discarding redundant, irrelevant and noisy features from the original set. This will not only help to increase the performance of the classifier but will also decrease the computational time

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 789–797.
springerlink.com

required to train the model. Feature selection is relevant in case of classification of microarray dataset considering the fact that among thousands of genes monitored simultaneously, only a fraction of them are biologically relevant. Therefore efficient feature selection methods are required to discover a set of discriminatory genes for effective class prediction and better clinical diagnose.

In literature, various feature selection tasks are classified into two categories [2]: filter method and wrapper method. In filter method, the feature subset selection is independent of the classifiers. Each feature subset is evaluated in terms of one of the class separability measures that exploit statistical properties of data for feature selection. Filter method requires less computation. Since the filter approach does not consider the learning bias introduced by final learning algorithm, it may not select the most relevant set of features for the learning algorithm. Also the selected feature set may contain correlated features in case of ranking approach which may degrade the performance of classifier. On the other hand, wrapper methods directly use the classification accuracy of some classifier as the evaluation criteria. They tend to find features better suited to the learning algorithm resulting in better performance. However, it is computationally more expensive since the classifier must be trained for each candidate subset. The conventional wrapper methods have been applied for feature selection on small or middle scale datasets. But, due to large computation time, it is difficult to apply them directly on high dimensional datasets. The computation time of wrapper methods can be reduced by reducing the search space. This is achieved by selecting a set of non-redundant features from the original set of features without losing any informative feature followed by a wrapper approach.

In this paper, a two stage approach is employed to determine a subset containing relevant and non redundant genes for better classification of microarray data. The first stage is to group correlated genes and then select one representative gene from each group to reduce redundancy. This requires partitioning of original gene set into some distinct clusters so that within a cluster genes are more similar or correlated while those in different clusters are dissimilar. In the second stage, a Sequential Forward Feature Selection method is applied to the set of genes obtained in the first stage to obtain a smaller set of discriminatory genes which can provide maximum classification accuracy. To remove redundancy, one way is to partition the set of genes such that within a cluster genes are more similar or correlated. Several clustering approaches [3, 4] have been proposed in literature such as partitioning method, kernel method, graph theoretic approach, hierarchical approaches and so on. We investigate four different clustering methods to cluster genes: k-means clustering, SOM clustering, hierarchical agglomerative clustering with complete linkage and hierarchical divisive clustering. Four publicly available and challenging cancer microarray datasets have been used to judge the performance of gene clustering methods in terms of classification accuracy and number of genes.

This paper is organized as follows. Next section includes brief description of state-of-art of gene clustering and the clustering methods used in this comparative study. Experimental setup and results are discussed in Section 3. Finally conclusions and future directions are included in last section.

2 Clustering Techniques for Gene Selection

In order to obtain better classification of microarray data, a smaller set of discriminatory genes needs to be identified. In literature researchers have developed various methods for gene selection. Some methods are based on gene scoring function which approximates the relative strengths of genes. Among them Golub et al.[5] used correlation measure which assumes that a discriminatory gene must have close expression levels in samples within a class, but significantly different expression levels in samples across different classes. Another model- free method with the assumption that discriminatory genes have different means across different classes, small intraclass variations and relatively large interclass variations is also suggested [6] for gene selection. Similarly, some other gene ranking approaches are suggested in literature [7].The problem with such ranking methods is that the gene subset returned may contain many correlated genes. Few wrapper based approaches [8] are also suggested for gene selection but due to high computational complexity are only suited to small and middle dimensional dataset. The computation time of wrapper method can be decreased by applying wrapper approach on only a smaller set of non-redundant genes.

In literature, clustering has been employed to obtain non-redundant genes. Clustering is the task of grouping objects according to some similarity measure, so that objects within same group are more similar compared to objects in other groups. Therefore genes can be clustered into groups to identify redundant and correlated genes. Many clustering approaches such as partition based, kernel based, graph theoretic, hierarchical approaches have been proposed in literature [3, 4] and still being developed to achieve better accuracy, stability and robustness. Most widely used clustering methods are K-means, hierarchical clustering, model based clustering, SOM, tight clusters etc. Each approach is associated with some advantages and disadvantages. In recent years gene clustering has gained much importance. Many researchers have proposed new methods or analyzed existing ones using different datasets. Tseng et al. [9] have used mouse embryonic data to compare different clustering methods for gene selection using rand index. Au et al. [10] have used attribute interdependence for clustering of genes and built a classifier by selecting significant genes from all groups. Cai et al. [11] have proposed clustered gene selection method to overcome the dimensionality problem. Mukhopadhayay et al. [12] have selected informative genes for Brain tumor and lung tumor datasets using multi-objective optimization clustering. However, to our best of knowledge, the comparison of different clustering methods to obtain correlated genes is not investigated. The four clustering approaches used to investigate clustering of correlated genes in this paper are briefly described below.

2.1 K-Means Method

The K-means algorithm [13] is a well known partition-based clustering algorithm which is simple and fast. It partitions the dataset into K (prespecified number) clusters by minimizing sum of squared distances between objects and their respective cluster centers. The objective function used in K-means is given by

$$D = \sum_{i=1}^{K} \sum_{x \in C_i} \left| x - \mu_i \right|^2 \tag{1}$$

Where x is a data object in cluster C_i and μ_i is the centroid of cluster C_i.

K-means algorithm generally converges in small number of iterations but is sensitive to noise and outliers. A global maximum is never assured while optimizing the objective function. Moreover the result is dependent on number of clusters K which is rarely known in advance. Choice of initial cluster centers also affects the clustering results. K-means algorithm generally returns clusters of hyper-spherical in shape.

2.2 Hierarchical Clustering Method

In contrast to K-means, Hierarchical Clustering [14] does not require user to specify the number of cluster, K. Since it generates a tree structure called dendrogram which is hierarchical series of nested clusters. The root node represents the complete dataset and every leaf is a data object. The tree can be cut at any level to obtain desired number of clusters. The algorithm gives clear visibility of hierarchical relationships among data objects and flexibility to choose desired number of clusters. Common disadvantage of hierarchical clustering is that it is sensitive to noise and outliers. Also it is not capable of correcting any misclassification errors since each object is considered exactly once. Cost of algorithm is high hence applicability is limited for large scale data. There are two approaches to hierarchical clustering: Agglomerative and Divisive.

2.2.1 Hierarchical Agglomerative Clustering

It is a bottom up approach where each object is considered as a separate cluster. At each step closest clusters are merged together until we are left with a single cluster. Different type of proximity measures can be used to merge clusters e.g. single linkage is a nearest neighbor based approach that considers the minimum of the distance between nearest points of two different clusters. Similarly, complete link is based on the farthest neighbor i.e. minimum distance between farthest points of two different clusters is determined for merging of clusters.

2.2.2 Divisive Hierarchical Clustering

Divisive clustering proceeds in a top down manner i.e. it starts with a single cluster containing all the data and then splits the clusters until there are only singleton clusters. Several heuristic methods have been proposed to decide splitting. We have used representative entropy [15,16] as splitting criteria which is given by

$$H_R = -\sum_{l=1}^{p} \overline{\lambda_l} \log \overline{\lambda_l} \tag{2}$$

Where $\overline{\lambda_l} = \dfrac{\lambda_l}{\sum_{l=1}^{p} \lambda_l}$ and $\lambda_l, l = 1 \ldots p$ are p eigen values of covariance matrix Σ of a cluster containing p genes.

High value of representative entropy signifies low redundancy in the cluster i.e. objects/genes are more dissimilar. Therefore the cluster with maximum entropy (low redundancy) is partitioned first. The process may be repeated to get desired number of clusters.

2.3 Self-Organizing Map (SOM)

SOM [17] is relatively more robust approach in case of noisy data. It generates a two dimensional map of given high dimensional data in order to place similar clusters near each other. The data objects are presented as neurons and adjacent neurons are connected to each other. Input patterns are fully connected to all neurons via adaptable weights, and during the training process, neighboring input patterns are projected into the lattice, corresponding to adjacent neurons. Thus it visualizes the latent structure of data. It requires users to specify the number of clusters desired which is difficult to specify in advance. Trained SOM may suffer from input space density misrepresentation, that is areas of low pattern density may be over-represented and areas of high density under-represented. Success of algorithm is dependent on proper choice of initial centers. The algorithm is more likely to determine hyper-spherical clusters.

3 Experimental Setup and Results

To investigate the performance of classification of microarray dataset on the choice of a clustering technique and a classification method, we have used four different clustering techniques: K-means, hierarchical agglomerative clustering, hierarchical divisive clustering, SOM; and three commonly used classification methods: K-Nearest Neighbor (KNN), Linear Discriminant Classifier (LDC) and Support Vector Machine (SVM). For evaluation purpose we have used four publicly available microarray datasets, which are considered to be challenging for classification. Three datasets are multiclass and one dataset is two class. Brief description of datasets used in this experiment is given in Table 1. The performance is evaluated in terms of classification accuracy and number of relevant genes. The datasets are preprocessed as described in [6] and then normalized using z-score before carrying out the experiments.

Table 1. Datasets Used

Dataset	Samples	Genes	Classes
CAR[18]	174	9182	11
GCM[19]	198	11328	14
NCI 60[20]	60	2000	8
Colon[21]	62	2000	2

Every dataset is clustered into sixty clusters by a clustering method. Representative gene from each cluster is selected using t-statistics. Thus we obtained a pool of 60 genes from each method for a given dataset. Thereafter, a sequential forward feature selection is applied to get a suboptimal set of genes which provide maximum classification accuracy.

The training and test data of CAR and GCM are already separately available. Hence, the classification accuracy of CAR and GCM are reported using test data. The classification accuracy of remaining two datasets is given in terms of both LOOCV and 10-fold cross validation. Results obtained are described in Table 2 and Table 3.

The number within parenthesis represents the number of genes selected for that dataset resulting in maximum classification accuracy for a given combination of clustering and classification method. Figure 1(a) shows the variation of average classification accuracy over all clustering methods with classification method for all datasets. Similarly, Figure 1(b) shows the variation of average classification accuracy over all classification methods with the choice of clustering method for all datasets. In case of 10-fold cross-validation, standard deviation over 50 repetitions is also specified. The following observations can be made from Tables 2-3 and Figures 1(a)-(b):

1. For CAR dataset maximum classification accuracy of 98.65% with 23 genes is achieved for hierarchical agglomerative clustering and LDC. Although same classification accuracy of 98.65% is also achieved with KNN classifier but with more number of genes. Average classification accuracy is maximum for hierarchical agglomerative clustering and KNN.
2. For GCM dataset a combination of hierarchical agglomerative clustering and SVM resulted in maximum accuracy 78.26% with 31 genes. Same classification accuracy is achieved with 37 genes with hierarchical agglomerative clustering and LDC. The overall performance measured in terms of average classification accuracy is maximum in case of hierarchical agglomerative clustering among four clustering methods and for KNN among the three classifiers used.

Table 2. Classification accuracies of microarray datasets

Dataset		Hier divisive	Hier agg.	K-means	Som
CAR	KNN	91.89(50)	98.65(48)	94.6(43)	97.3(16)
	LDC	91.89(50)	98.65(23)	91.89(46)	95.95(33)
	SVM	93.24(28)	94.6(26)	95.95(36)	95.95(33)
GCM	KNN	69.57(26)	76.09(13)	69.57(30)	67.39(26)
	LDC	63.04(16)	78.26(37)	65.22(12)	65.22(19)
	SVM	63.04(45)	78.26(31)	71.74(31)	67.39(35)
Nci60	KNN	83.33(34)	90(26)	91.67(40)	88.33(16)
	LDC	89.66(14)	93.1(13)	82.76(13)	89.66(15)
	SVM	88.33(15)	86.67(17)	94.83(14)	91.67(16)
Colon	KNN	95.16(14)	96.77(24)	96.77(5)	95.16(12)
	LDC	93.55(36)	91.94(16)	91.94(10)	90.32(7)
	SVM	93.55(4)	93.55(7)	96.77(6)	95.16(14)

3. For NCI60 dataset maximum classification accuracy of 94.83% with 14 genes is achieved for K-means clustering and SVM. Average classification accuracy is maximum for hierarchical agglomerative clustering and SVM.
4. For colon dataset two combinations K-means and KNN & K-means and SVM resulted in maximum accuracy 96.77% with 5 genes and 6 genes respectively. Although same classification accuracy of 96.77% is also achieved with hierarchical agglomerative method but with more number of genes. The overall performance measured in

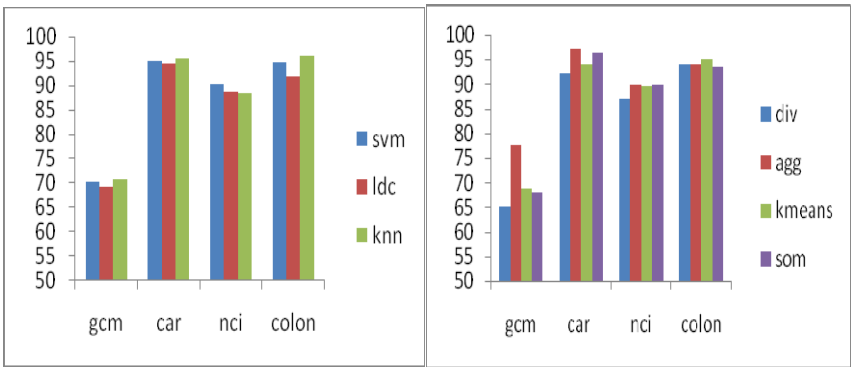
terms of average classification accuracy is maximum in case of K-means among four clustering methods and for KNN among the three classifiers used.

5. The variation in classification accuracy between 10 fold cross-validation and LOOCV is not significant for colon dataset. However the maximum variation of 10% is observed for some combinations of clustering and classification methods for multiclass NCI60 dataset.

Table 3. Classification accuracies of 10-fold cross-validation

Dataset		Hier divisive	Hier agg.	K-means	Som
NCI60	KNN	84.24(21)±1.81	89.1(27) ±2.19	91.55(31)±2.21	80.07(31)±2.61
	LDC	89.48(27)±2.52	93.1(29)±2.12	88.38(18)±2.48	88.86(22)±2.04
	SVM	84.03(12)±3.26	92.03(34)±2.39	92.34(32)±2.49	85.65(29)±1.95
Colon	KNN	94.48(39)±0.98	96.39(11)±0.70	96.42(10)±0.82	93.55(24) ±0.0
	LDC	93.52(25)±2.15	92.90(35)±2.75	91.55(10)±0.90	91.48(15) ±0.86
	SVM	93.52(18)±0.23	96.77(16) ±0.00	93.55(8) ±0.00	93.52(11) ±0.23

6. The performance of Hierarchical agglomerative method is best among all analyzed methods. It resulted in maximum average classification accuracy for three datasets i.e. CAR, GCM and NCI60.
7. Among the three classifiers analyzed, the performance of KNN is best which achieves maximum average classification accuracy for three datasets viz. GCM, CAR and Colon.



(a) Averaged over four clustering methods (b) Averaged over three classifiers

Fig. 1. Performance in terms of average classification accuracy

In Table 4, the best performance of gene clustering is compared with already existing approaches. The classification accuracies obtained were comparable for GCM and Colon and considerably higher than other state of art methods for CAR and NCI datasets. For colon dataset although accuracy is a bit lower but number of genes is very small compared to existing approaches.

Table 4. Comparison of classification accuracy with other state of art methods

CAR		GCM		NCI		Colon	
Agg. +LDC	**98.65 (23)**	**Agg. +SVM**	**78.26 (31)**	**K-means +SVM**	**94.83 (14)**	K-means +KNN	96.77 (5)
GS1 +SVM[6]	90.2 (100)	Zhang et al[22]	64.65	Zhang et al [22]	68.33	Pso+ann[6]	88.7
Cho et al.[24]	87.9 (97)	Wang et al.[23]	69.8 (28)	WFF SA-G* [25]	85.25 (14)	WFFSA-G* [25]	97.9 (100)
F-test.[6]	88.5 (97)	OVA-SVM[19]	78	mRMR_d +KNN[25]	89.66 (95)	**NCUT+LDC [14]**	**98.38 (32)**

4 Conclusion

In literature, the conventional wrapper methods have been applied for relevant feature selection on small or middle scale datasets. However, it is difficult to apply directly on high dimensional microarray datasets due to large computation time. In this paper, a two stage approach is proposed to determine a subset containing relevant and non redundant genes for better classification of microarray data. In first stage, genes were partitioned into distinct clusters so that genes within a cluster are highly correlated. To determine the better choice of clustering algorithm for finding relevant and non redundant genes, four different clustering methods were investigated. Experiments on four well known cancer microarray datasets depicted that hierarchical agglomerative with complete link approach performed best in terms of average classification accuracy for three datasets: CAR, GCM and NCI60. For colon dataset, K-means performed better than other approaches. Also KNN emerged out to be the best classifier for GCM, CAR and colon datasets whereas SVM performed best for NCI60. It is also noted that variation in 10 fold and LOOCV accuracies is not significant for two class colon dataset but is considerable for multiclass NCI60 dataset. The difference may be attributed to class imbalance present in NCI60 dataset. Comparison with other state of art methods have shown that the proposed approach which involves gene clustering is effective in reducing redundancy among selected genes to provide better classification of microarray datasets.

References

1. Bellman, R.: Adaptive Control Processes. In: A Guided Tour. Princeton University Press (1961)
2. Guyon, I., Elisseeff, A.: An Introduction to Variable and feature Selection. Journal of Machine Learning Research (3), 1157–1182 (2003)
3. Jain, A.K., Murthy, M.K., Flynn, P.J.: Data Clustering: A Review ACM Computing surveys, vol. 31, pp. 264–318
4. Rui, X., Wunsch, D.: Survey of Clustering Algorithms. IEEE Transactions on Neural Networks 16(3) (2005)

5. Golub, T.R., Slonim, D.K., Tamayo, P., et al.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286, 531–537 (1999)
6. Yang, K., Cai, Z., Li, J., Lin, G.H.: A stable gene selection in microarray data analysis. BMC Bioinformatics (2006) 1471-2105-7-228
7. Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97(457), 77–87 (2002)
8. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence 97(1-2), 273–324 (1997)
9. Tseng, G.C., Wong, W.H.: Tight Clustering: A Resampling-based Approach for Identifying Stable and Tight Patterns in Data Biometrics, vol. 61, pp. 10–16 (2005)
10. Au, W.H., Chan, K.C., Wong, A.K., Wang, Y.: Attribute Clustering for Grouping, Selection, and Classification of Gene Expression Data. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 2(2), 83–101 (2005)
11. Cai, Z., Xu, L., et al.: Using clustering to identify discriminatory genes with higher classification accuracy. In: IEEE Symposium 0-7695-2727-2/06
12. Mukhopadhyay, A., et al.: Simultaneous Informative Gene Selection and Clustering through Multiobjective Optimization. IEEE Congress on Evol. Comp., 1–8 (2010)
13. Tavazoie, S., Huges, D., Campbell, M.J., Cho, R.J., Church, G.M.: Systematic determination of genetic network architecture. Nature Genet., 281–285 (1999)
14. Eisen, M.B., Spellman, T.P., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. USA 95(25), 14863–14868 (1998)
15. Pal, S.K., Mitra, P.: Pattern recognition algorithms for data mining. Chap. and Hall (2008)
16. Bala, R., Agrawal, R.K., Sardana, M.: Relevant Gene Selection Using Normalized Cut Clustering with Maximal Compression Similarity Measure. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds.) PAKDD 2010, Part II. LNCS, vol. 6119, pp. 81–88. Springer, Heidelberg (2010)
17. Kohonen, T.: Self-organizing maps. Springer, Berlin (1995)
18. Su, A.I., Welsh, J.B.: Molecular classification of human carcinomas by gene expression signatures. Cancer Research 61, 7388–7393 (2001)
19. Ramaswamy, S., Tamayo, P., Rifkin, R., et al.: Multi-class cancer diagnosis using tumor gene expression signatures. PNAS 98, 15149–15154 (2001), Dataset description
20. Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., et al.: Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines. Nature Genet. 24, 227–235 (2000)
21. Alon, U., Barkai, N., et al.: Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Array. Proc. Nat'l Academy of Science 96(12), 6745–6750 (1999)
22. Zhang, Y., Ding, C., Li, T.: Gene selection algorithm by combining Relief and RMR. BMC Genomics 9(2), S27 (2008)
23. Wang, L., Chu, F., Xie, W.: Accurate cancer classification using expressions of very few genes. IEEE/ACM Trans. on Comp. Biology and Bioinformatics 4(1) (2007)
24. Cho, J., Lee, D., Park, J.H., Lee, I.B.: New gene selection for classification of cancer subtype considering within-class variation. FEBS Letters 551, 3–7 (2003)
25. Zhu, Z., Ong, Y.-S., Monoranjan, D.: Wrapper–Filter Feature Selection Algorithm Using a Memetic Framework. IEEE Trans. Cybernatics 37(1) (2007)

An Optimal Approach for DICOM Image Segmentation Based on Fuzzy Techniques

J. Umamaheswari[1] and G. Radhamani[2]

[1] Research Scholar and [2] Director, Department of Computer Science,
Dr. G.R.D. College of Science,
Coimbatore, Tamilnadu, India
umamugesh@yahoo.com, radhamanig@hotmail.com

Abstract. In this paper an optimal method for DICOM CT image segmentation is explored with the integration of FCM thresholding with fuzzy levelset for medical image processing FCM thresholding gives fine segmented results when compared to Otsu method. The optimization property of FCM is improved when it is combined with local thresholding. The application of Fuzzy levelset gives enhanced segmentation results. The experimentation results based on the statistical metrices proves that the optimal approach enhances the segmented results with fine regions.

Keywords: Medical Image segmentation, FCM, Local thresholding, Levelset, Fuzzy levelset, Adaptive thresholding.

1 Introduction

Segmentation of images holds an important position in the area of image processing and widely used in medical applications which includes surgical planning, abnormality detection and treatment progress monitoring. The purpose of segmentation is to partition an image into distinct, semantically meaningful entities by defining boundaries between features and objects in an image based on some constraint, or homogeneity predicate. Computer aided detection of abnormal growth of tissues is primarily motivated by the necessity of achieving maximum possible accuracy. The various hybrid models for medical image segmentation is explained in [1,4].

There are many methods available using levelset segmentation [9,10,13,14]. Such algorithms employ fuzzy clustering, based on image intensity, for initial segmentation and employ levelset methods for object refinement by tracking boundary variation. The previous work on liver tumor segmentation [9] has based on fuzzy clustering. These methods approximately delineate a tumor boundary which not only relieves manual intervention, but also accelerates levelset optimization. Ho and Suri, on the other hand, proposed to regularize levelset evolution locally by fuzzy clustering, in order to deviate the problems of noises sensitivity and weak boundaries of medical images [10,13,14].

An optimized approach that uses a variational method of image segmentation is described to minimize energy by weighted Total Variation (TV) method is discussed in [2,3]. This model describes the Fuzzy C-means method (FCM) to obtain segmentation

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 799–807.
springerlink.com

via fuzzy pixel classification. FCM allows pixels to fit in multiple classes with varying degrees of membership than hard classification methods that normally makes pixel to fit in one class. This approach allows additional flexibility in many applications and has recently been used in processing of magnetic resonance image (MRI) and Computer Topography (CT) image. Threshold segmentation is widely used in many fields because of its simplicity and efficiency which is most frequently used for image segmentation. In our work the integration of FCM thresholding with fuzzy levelset is analyzed. The segmented results are proved to be efficient compare to other traditional method.

The paper is organized as follows, section 2 discusses about the optimal method for DICOM image segmentation. Section 3 describes the experimentation and results. Finally the section 4 includes the conclusion and references.

2 Overview of the Optimal Approach for DICOM Image Segmentation

The optimal approach consists of FCM thresholding with fuzzy levelset process. The FCM algorithm that incorporates spatial information into the membership function is used for clustering, while a conventional FCM algorithm does not fully utilize the spatial information in the image. The advantages of the algorithm are its less sensitivity to noise and consider regions more homogeneous compared to other methods. So fuzzy levelset by Bing Nang Li [15] is used for the enhancement of the segmentation results. The optimal approach gives better results for DICOM image segmentation.

2.1 FCM Thresholding

The fuzzy c means clustering method with thresholding approach is based on the principles of fuzzy algorithm. It consists of three components namely, Fuzzy clustering, C -Means and Thresholding.

a. Fuzzy Clustering

The goal of a clustering analysis is to divide a given set of image data or objects into a cluster, which represents subsets or a group. The partition should have two properties such as homogeneity and heterogeneity. The homogeneity cluster consists of similar data whereas the heterogeneity cluster contains different data. The membership functions do not reflect the actual data distribution in the input and the output spaces. They may not be suitable for fuzzy pattern recognition. To build membership functions from the data available, a clustering technique may be used to partition the data, and then produce membership functions from the resulting clustering. In our method C-means clustering is used. It is a simple unsupervised learning method which can be used for data grouping or classification when the number of the clusters is known. It consists of the following steps:

Step 1: Choose the number of clusters K
Step 2: Set initial centers of clusters C1, C2,…, C.

Step 3: Classify each vector $x_l = [x_{l1}, x_{l2}, \ldots x_{ln}]^T$ *into the closest center* c_i *using Euclidean distance measure:* $\|x_i - c_i\| = \min\|x_i - c_i\|$

Step 4: Recompute the estimates for the cluster centers c_i.

Let $c_i = [c_{l1}, c_{l2}, \ldots c_{ln}]^T$ c_{im} *be computed by:* $c_{im} = \frac{\sum x_{li} \in cluster(i^{x\lim})}{N_i}$

where N_i *is the number of vectors in the i-th cluster.*

Step 5: If none of the cluster centers (c_i *=1, 2,..., k) changes in step 4 stop process, otherwise go to step 3.*

b. C-means Algorithm: The criterion function used for the clustering process is:

$$J(v) = \sum_{k=1}^{n} \Sigma k_x \in c_i \left| x_k - v_i \right|^2,$$

where v_i is the sample mean or the center of samples of cluster *i*, and $v=\{v1,v2,\ldots,vc\}$.

Clusters are not completely disjoint and the data could be classified as belonging to one cluster almost as well to another. Therefore, the separation of the clusters becomes a fuzzy notion, and representation of the data can be more accurately handled by fuzzy clustering methods. It is necessary to describe the data in terms of fuzzy clusters. The criterion function used for fuzzy C-means clustering is

$$J(v) = \sum_{i=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} \left| x_k - v_i \right|^2,$$

where:

$x_1, \ldots\ldots x_n$ – *'n'* data sample vectors;

$v_1, \ldots\ldots v_c$ – *'c'* denotes cluster centers (centroids);

$u = u_{ik}$ *cxm* matrix, where u_{ik} is the *i*-th membership value of the *k*-th input sample x_k, and the membership values satisfy the following conditions:

$$0 \le U_{ik} < 1; i = 1, \ldots\ldots c; k = 1, \ldots\ldots, n;$$

$$\sum_{i=1}^{C} U_{ik} = 1; k = 1, \ldots\ldots n;$$

$$0 < \sum_{k-1}^{n} U_{ik} < 1; i = 1, \ldots\ldots c;$$

$m \in [1, \infty)$ is an exponent weight factor.

c. Thresholding

Local threshold method is used to automatically perform histogram shape-based image thresholding or, the reduction of a graylevel image to a binary image. The algorithm assumes that the image to be threshold contains two classes of pixels or bimodal histogram (e.g. foreground and background) then calculates the optimum threshold separating those two classes so that their combined spread is minimal. Through FCM method the segmented part cannot be seen visibly. Here FCM is used based on local thresholding. The segmentation image is made visible through optimized method.

2.2 Fuzzy Levelset for Image Segmentation

This approach is based on the active [5-8] fuzzy model with the integration of adaptive region information to obtain a robust segmentation model. The level set method was first introduced by Osher and Sethian [17]. The level set method is a numerical and theoretical tool for propagating interfaces. Both FCM algorithms and level set methods are general-purpose computational models that can be applied to problems of any dimension. However, when applied to medical image segmentation, it helps to consider the specific circumstances for better performance. A new fuzzy level set algorithm is thereby proposed for automated medical image segmentation [15]. It begins with spatial fuzzy clustering, whose results are utilized to initiate level set segmentation, estimate controlling parameters and regularize level set evolution. It employs an FCM with spatial restrictions to determine the approximate contours of interest in a medical image. Benefitting from the flexible initialization as in levelset, the enhanced level set function can accommodate FCM results directly for evolution.

Suppose the component of interest in an FCM results, the level set function φ of the closed front C is defined as follows, [16]

$$\varphi(x, y) = \pm d\left((x, y), C\right)$$

Where $d((x, y),C)$ is the distance from point (x, y) to the contour C, and the sign plus or minus are chosen if the point (x, y) is inside or outside of interface C. The interface is now represented implicitly as the zero level set (or contour) of this scalar function

$$C = \left\{(x, y) / \varphi(x, y)\right\}$$

Such an implicit representation has numerous advantages over a parametrical approach. The level set evolution equation is given by

$$\frac{\partial \varphi(x, y)}{\partial t} = \delta_c(\varphi(x, y))\left[\begin{array}{l} \nu k(\varphi(x, y)) \\ \left\{\left(I(x, y) - \mu_1\right)^2 - \left(I(x, y) - \mu_0\right)^2\right\} \end{array}\right]$$

Where μ0 and μ1 are the mean of the image intensity within two subsets inside or outside the contours respectively. The final segmented image can be represented as a set of piece-wise constants.

2.3 Fuzzy Thresholding and Fuzzy Levelset Method

While applying normal thresholding to the edges, the region of the image is not segmented properly and hence the excess regions are removed by applying FCM method. An FCM algorithm which is the general-purpose computational model is applied to DICOM CT image segmentation, for the better results. An image can be represented in various feature spaces, and the FCM algorithm classifies the image by grouping similar data points in the feature space into clusters. The resultant image obtained by FCM consists of less spatial information. So to improve that and to separate the segmented part from the spatial region, the local thresholding is applied. The image obtained by fuzzy thresholding consist of some deblurred edges and some incorrect segmented part. To improve the segmentation the fuzzy levelset method is used. The levelset function will automatically slow the evolution down and will become totally dependent on the smoothing term. Since a conservative levelset evolution is adopted here, it stabilizes automatically. For robust segmentation a comparatively large iteration of evolution is adopted. The levelset evolution is applied in order to avoid insufficient or excessive segmentation.

3 Experimentation and Results

DICOM medical images are taken as test images for evaluating results. Here an average of ten images is taken for evaluation. The algorithm is tested in MATLAB. The reconstruction of an image has the dimensions of 256 pixel intensity. The DICOM CT images in this contain a wide variety of subject matters and textures. Most of the images used are brain images with normal and abnormal images. The results are estimated statistically based on The Energy, Entropy, UQI (Universal Quality Index) and Mutual Information (MI)

To test the accuracy of the segmentation algorithms, four steps are followed.

i) First, a DCIOM CT image is taken as test input.
ii) Second, the different segmentation method is applied to a DICOM image.
iii) Third, the performance evaluation is obtained based on the statistical measures like Energy, Entropy, UQI and MI.

a) Energy: The gray level energy indicates how the gray levels are distributed. It is formulated as,

$$E(x) = \sum_{i=1}^{x} p(x)$$

where E(x) represents the gray level energy with 256 bins and p(i) refers to the probability distribution functions, which contains the histogram counts. The larger energy value corresponds to the lower number of gray levels, which means simple. The smaller energy corresponds to the higher number of gray levels, which means complex. The following table 1 shows the different parametric evaluation for segmentation algorithm.

Table 1. Parametric Evaluation for Different Segmentation algorithms

Method	Energy	Entropy	UQI	MI
Adaptive threshold	5.90E-01	7.06E-01	0.1046	4.38E-01
Otsu Threshold	5.94E-01	8.59E-01	0.10708	8.59E-01
Fuzzy Threshold	5.50E-01	9.27E-01	0.1571	7.38E-01
Fuzzy Levelset	5.65E-01	6.25E-01	0.1892	6.24E-01
HM	6.90E-01	1.48E+00	0.3542	1.43E+00

b) Entropy: Suppose that two discrete probability distributions of the images have the probability functions of p and q, the relative entropy of p with respect to q is then defined as the summation of all possible states of the system, which is formulated as,

$$d = \sum_{i=1}^{k} p(i) \log_2 \frac{p(i)}{q(i)}$$

The following figure 1 shows the energy value for segmentation algorithm. The hybrid method gives high energy value compare to other methods.

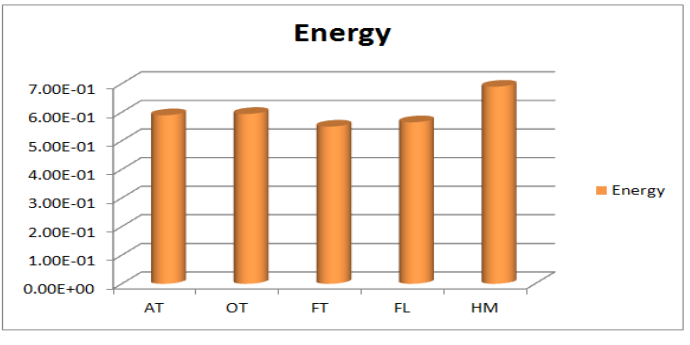


Fig. 1. Energy Value of Different Segmentation Algorithm

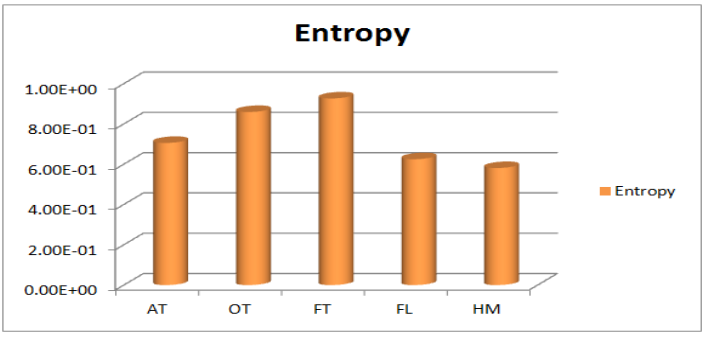


Fig. 2. Entropy Value for Different Segmentation Algorithm

The above figure 2 shows the entropy value for segmentation algorithm. The hybrid method gives less entropy value compared to other methods.

c) UQI: UQI measures image similarity across distortion types. Distortions in UQI are measured as a combination of three factors; Loss of correlation, Luminance distortion and Contrast distortion. Let $\{x_i\}$ and $\{y_i\}$ =1,2,...,N be the original and the test image signals, respectively. The universal quality index is defined as

$$UQI = \frac{4\sigma_{xy}\overline{x}\,\overline{y}}{\left[\sigma_x^2 + \sigma_x^2\right]\left[\left(\overline{x}\right)^2 + \left(\overline{y}\right)^2\right]}$$

The following figure 3 shows the UQI value for segmentation algorithm. The hybrid method gives high UQI value compared to other methods.

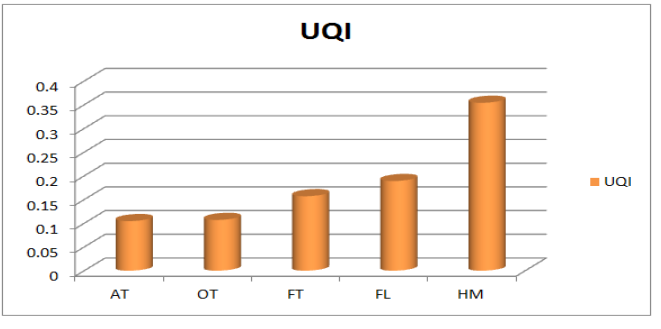


Fig. 3. UQI Value for Different Segmentation Algorithm

d) Mutual Information (MI): The notion of the mutual information can be applied as another objective metric. The mutual information acts as a symmetric function, which is formulated as,

$$I(X,Y) = \sum_{XY} P_{XY}(X,Y) \log_2 \frac{P_{xy}(X,Y)}{P_x(X)P_y(Y)}$$

$$= -\sum_x P_x(X) \log_2 P(X) + \sum_{x,y} P_{xy}(X,Y) \log_2 \frac{P_{xy}(X,Y)}{P_x(X)P_y Y)}$$

$$= H(X,) - H(X \mid Y)$$

where I(X; Y) represents the mutual information; H(X) and H(X|Y) are entropy and conditional entropy values. It is interpreted as the information that Y can tell about X and the measure of reduction in uncertainty of X due to the existence of Y. At the same time, it also shows the relationship of the joint and product distributions. The following figure 4 shows the MI value for segmentation algorithm. The hybrid method gives high MI value compared to other methods.

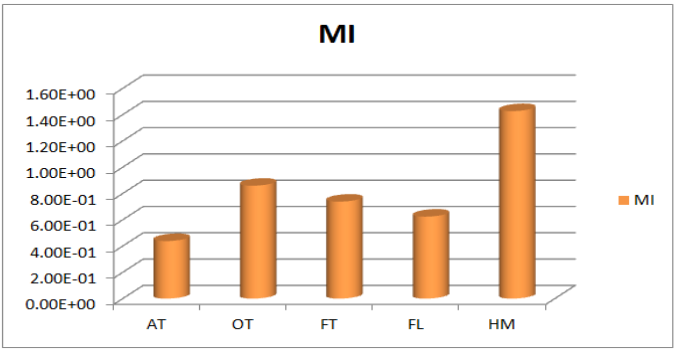


Fig. 4. MI Value for Different Segmentation Algorithm

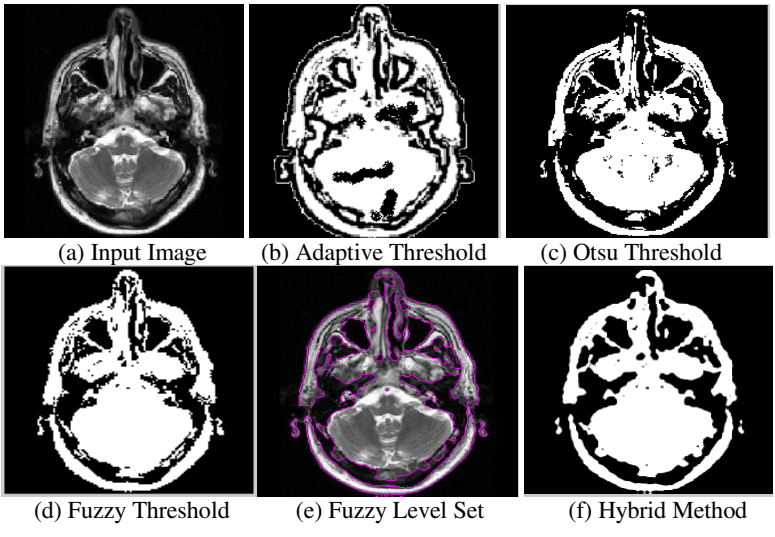


Fig. 5. Image Results for Different Segmentation Algorithm

The above figure 5 shows the image results for different segmentation algorithm. The hybrid method gives the suitable segmented results with fine regions.

4 Conclusion

In this paper an optimal method for DICOM CT image segmentation is explored with the integration of FCM thresholding with fuzzy levelset for medical image processing FCM thresholding gives fine segmented results when compared to Otsu method. The optimization property of FCM is improved when it is combined with local thresholding. The application of Fuzzy levelset gives enhanced segmentation results. The experimentation results based on the statistical metrices proves that the optimal approach enhances the segmented results with fine regions.

References

1. Yang, F., Xiaohuan: An Improved Hybrid Model for Medical Image Segmentation. In: Proceeding of IEEE Conference, ICCS 2008, pp. 367–370 (2008)
2. Soesanti, I., Susanto, A., Widodo, T.S., Tjokronagoro, M.: Optimized Fuzzy Logic Application for MRI Brain Images Segmentation. International Journal of Computer Science & Information Technology (IJCSIT) 3(5), 137–146 (2011)
3. Soesanti, I., Susanto, A., Widodo, T.S., Tjokronagoro, M.: MRI Brain Images Segmentation Based on Optimized Fuzzy Logic and Spatial Information. International Journal of Video & Image Processing and Network Security IJVIPNS-IJENS 11(04), 6–10 (2011)
4. Zulaikha Beevi, S., Mohamed Sathik, M.: An Effective Approach for Segmentation of MRI Images: Combining Spatial Information with Fuzzy C-Means Clustering. European Journal of Scientific Research 41(3), 437–451 (2010)
5. Ghassabeh, Y.A., Forghani, N., Forouzanfar, M., Teshnehlab, M.: MRI Fuzzy Segmentation of Brain Tissue Using IFCM Algorithm with Genetic Algorithm Optimization. In: Proceeding of IEEE International Conference, pp. 665–668 (2007)
6. Senthilkumaran, N., Rajesh, R.: Brain Image Segmentation. International Journal of Wisdom Based Computing 1(3), 14–18 (2011)
7. Ibrahim, S., Khalid, N.E.A., Manaf, M.: Seed-Based Region Growing (SBRG) vs Adaptive Network-Based Inference System (ANFIS) vs Fuzzy C-Means (FCM): Brain Abnormalities Segmentation. International Journal of Electrical and Computer Engineering 5(2), 94–104 (2010)
8. Wiselin Jiji, G., Ganesan, L.: Unsupervised Segmentation using Fuzzy Logic based Texture Spectrum for MRI Brain Images. World Academy of Science, Engineering and Technology 5, 155–157 (2005)
9. Zhang, J., Hu, J.: Image Segmentation Based on 2D Otsu Method with Histogram Analysis. In: 2008 International Conference on Computer Science and Software Engineering, pp. 105–108. IEEE (2008)
10. Suri, J.S., Liu, K., Singh, S., Laxminarayan, S.N., Zeng, X., Reden, L.: Shape recovery algorithms using levelset sin 2-D/3-D medical imagery: a state-of-the-art review. IEEE Transactions on Information Technology in Biomedicine, 8–28 (2002)
11. Paragios, N.: A levelset approach for shape-driven segmentation and tracking of left ventricle. IEEE Transactions on Medical Imaging, 773–776 (2003)
12. Mitchell, I.M.: The flexible, extensible and efficient toolbox of levelset methods. Journal of Scientific Computing, 300–339 (2008)
13. Suri, J.S.: Two-dimensional fast magnetic resonance brain segmentation. IEEE Engineering in Medicine and Biology Magazine 20, 84–95 (2001)
14. Ho, S., Bullitt, E., Gerig, G.: Levelset evolution with region competition: automatic 3-D segmentation of brain tumors. In: Proceedings of the International Conference on Pattern Recognition (ICPR 2002), pp. 532–535 (2001)
15. Bing, N., Chee, K., Chang, S., Ong, S.H.: Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation, vol. 41, pp. 1–10 (2011)
16. Li, B.N., Chui, C.K., Ong, S.H., Chang, S.: Integrating FCM and levelsets for liver tumor segmentation. In: Proceedings of the 13th International Conference on Biomedical Engineering (ICBME 2008), pp. 202–205 (2009)
17. Sethian, J.A., Smereka, P.: Levelset Methods for Fluid Interfaces. Annu. Rev. Fluid Mech. 35, 341–372 (2003)

A Two-Phase Item Assigning in Adaptive Testing Using Norm Referencing and Bayesian Classification

R. Kavitha[1], A. Vijaya[2], and D. Saraswathi[3]

[1] Assistant Professor, Dept. of MCA, AIMIT, St. Aloysius College (Autonomous), Mangalore, Karnataka, India
kavitharajamanii@yahoo.co.in
[2] Assistant Professor, Dept. of Computer Science, Govt. Arts College (Autonomous), Salem, Tamilnadu, India
[3] Assistant Professor, Dept. of Computer Science, KSR College of Arts and Science, Tiruchengode, Tamilnadu, India

Abstract. Due to the advancement in information technology and varied learner group, e-learning has become popular. Hence computer based assessment become a prevalent method of administering the tests. Randomization of test items here may produce unfair effect on test takers which is unproductive in the outcome of the test. There is a need to develop the Intelligent Tutoring System that assigns intelligent question depending on the student's response in the testing session. It will be more productive when the questions are assigned based on the ability in the early stage itself. Also, if only the standard multiple-choice questions are focused, then the real embedded nature of computer assessment is sacrificed. Items with different constrained constructs are included to bring out the complex skills, analytical and comprehensive ability of learners. So, this study focus on building up a framework to automatically assign intelligent question with different constructs based on the learner ability while entry. Using Norm Referencing, questions are classified based on item difficulty. Item discrimination is found and there by only the items which can discriminate the performers alone are accumulated in the item pool to have maximum effect of intelligence in tutoring system. The level of new learner is predicted by means of Naïve Bayesian classification and the consequent item is posed. Thereby the objective of Intelligent Tutoring System is achieved by using both adaptability and intelligence in testing.

Keywords: Intelligent Item Classification, Adaptivity in ITS, Norm Referencing in ITS.

1 Introduction

Computers and electronic technology today offer myriad ways to enrich educational assessment both in the classroom and in large-scale testing situations. The fast growing popularity of the e-tests is because of the many advantages they have: creation with minimum efforts and using graphical interface; sharing and reusing of tests from different target groups; objectivity and automatic assessment of the questions; time saving,

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 809–816.
springerlink.com

etc.[1]. Through these and other technological innovations, the computer-based platform offers the potential for high quality formative assessment.

Each student has their own learning style, and this yields a result that each student's performance in learning cannot be assessed and evaluated in a unique and simple way. Also it cannot be evaluated only by measuring the test results depending on the number of right and wrong answers. Therefore, how to progress an efficient learning process is a critical issue. For this, testing must be intelligent. It must behave as if a teacher asks questions to a student in real class environment. If student cannot answer the question, teacher must ask easier question about similar subject and if student answer this time, then, again more difficult question must be asked.

Another critical issue is the question type currently dominating large-scale computer-based testing and many e-learning assessments is the standard multiple-choice question, which generally includes a prompt followed by a small set of responses from which students are expected to select the best choice. This kind of task is readily scorable by a variety of electronic means and offers some attractive features as an assessment format. However, if e-learning developers adopt this format alone as the focus of assessment formats in this emerging field, much of the computer platform's potential for rich and embedded assessment could be sacrificed. Thus, by combining intermediate constraint types and varying the response and media inclusion, e-learning instructional designers can create a vast array of innovation assessment approaches and could arguably match assessment needs and evidence for many instructional design objectives.

The regard is the testing part and the concern is to develop an intelligent testing application that will produce intelligent questions depending on the student's responses and performance during testing session. This kind of testing is called as Computer Adaptive Testing (CAT). CAT is a methodology of testing which adapts to the examinee's level. CAT selects questions in order to maximize the performance of examinee by observing the past success throughout the test. Therefore, the difficulty of test depends on the examinee's performance and level of ability. This is how the Intelligent Tutoring System (ITS) can be developed. It is a system that provides direct or indirect customized instruction or feedback to learners whilst performing a task. In this regard, the question levels must be determined somehow. In this study, not only item difficulties, but also item discrimination of questions was estimated using item responses by using Norm Reference Approach. The items which are not well discriminating between high, medium and low performers are discarded. At the same time, new learner ability can be predicted by learning through the Naïve Bayesian Classification approach. So, at the commencement of the test, the intelligent question is assigned to the learners based on their ability. Thereby the purpose of adaptivity in Intelligent Tutoring System testing environment is attained.

The next section explores the background study. The third section gives an insight on methodology proposed in this study with a detailed description flows involved in two phases. The final section gives the conclusion and further enhancements for this study.

2 Literature Survey

There are lots of academicals and commercial work done on computer based testing applications. The need of speed, time flexibility, low-cost, fair scoring and besides the

unceasingly increasing information technology makes the computer based testing applications essential.

In computer based tests, randomized presentation of items is automatically programmed into testing software to present different items to the test takers. The downside of such randomization is that it prevents planned sequencing of items. Randomizing items does not accommodate a test user or a constructor who wishes to ensure that items progressively become tougher. It may unfairly increase test anxiety for some of the candidates. Increased anxiety at any stage during the test for whatever reason is likely to have a negative effect on that person's performance for the remainder of the test [2]. In a research study [3], it was proved that randomization is ensuring the test security, yet progressively allowing items to become more difficult as the test items are presented to each test-taker, will prevent occurrence of the item randomization effect. Instead of giving each examinee the same fixed test, CAT item selection adapts to the ability level of individual examinees. After each response, the examinee's ability estimate is updated and the subsequent item is selected to have optimal properties at the new estimate. According to some researchers, ubiquitous multiple-choice testing sometimes encourages "poor attitudes toward learning and incorrect inferences about its purposes. For example that there is only one right answer, that the right answer resides in the head of the teacher or test maker, and that the job of the student is to get the answer by guessing" [4].

The development of item response theory (IRT) and Norm Referenced Test (NRT) in the middle of the last century has provided a sound psychometric footing for CAT. It is a modern test theory and is currently an area of active research. The key feature of NRT is its modeling of response behavior with distinct parameters for the examinee's ability and the characteristics of the items. The need to upgrade from ordinary CBT to CAT is well concentrated and illustrated [5]. Asking an easy question to a high ability student would not provide true information about his/her ability even the answer is correct. Likewise, a difficult question answered wrongly by a less successful student would not show the real ability level of the student. By selecting and administering questions that match the individual student's estimated level of ability, questions that present low value information can be avoided [6]. Low performance student might be disappointed and high performance students might be bored and tired of questions with inappropriate levels of difficulty. So it can be stated that in addition to increasing efficiency, CAT also increase the level of interaction and motivation of the student.

S.C. Cheng et al [7] proposed an automatic leveling system for e-learning examination pool using entropy measure. The questions were leveled based on the response given by the greater part of learners with similar background. In order to assess the capacity of each question or task to distinguish between those who know and those who do not, the trial group of candidates should possess a range of knowledge from those with good knowledge to those lacking it [8].

From the literature, it is very well seen that, there is a need of adaptive assessment with intelligence through some new enrichments.

3 Methodology

Assessment involves selecting evidence from which inferences can be made about current status in learning sequence. This kind of instructional methodology can develop education quality and efficiency. Since, the Computer Adaptive Testing pose questions based on the ability of learner, the item difficulty of the question has to be found out initially. All available tests in CAT are assigning the question with medium level difficulty to the learner first. Then based on the response, successive questions depending on their capability level are posed. This study involves finding the learning level of the learner first. Hence, their entry into the test itself is restricted based on individual ability. Traditional test analysis considers the extent to which a single item distinguishes between able and less able candidates in a similar way to the test as a whole. Items which are not consistent with the other items in the way in which they distinguish between able and less able candidates are considered for deletion. All the previous studies focused only on how well the item is discriminating high and low performers alone. Since, there will be more number of learners falls on the category of average learners, this study focus on how well the item discriminates Medium and Low performers, High and Medium performers. This conform the effectiveness of Intelligent Tutoring Systems.

When only the standard multiple-choice based items are focused, then the practical embedded assessment in adaptivity could be given up. Hence, to enhance the analytical thinking ability, comprehensive ability, various intermediate constraint constructs item are used. The Type1 belongs to standard True / False and Multiple-Choice based questions. Type2 includes Selection and Identification. Type3 focus on Reordering and Rearrangement and final Type4 on Substitution and Correction. Hence, the item types are gradually upgrading from fully constrained responses to constructed responses. Naïve Bayesian Classification is used to predict the level of a new learner and specific difficulty level of item is posed during the testing session.

The Intelligent Item Posing proposed here includes two phases. The first phase focuses on Accumulation of Labeled Items. Grade Fuzzification is used to transform the numeric grade into symbolic data with high, medium and low degree level of indication. The Item Difficulty and Item Discrimination are found using Item analysis by Norm Referencing approach. Those Items which have negative discrimination index are pruned. Because, those items are not well discriminates the High and Medium performers or Medium and Low performers. The reason may be wrong key answer for the item or it may not be well framed. These may lead to erratic analysis and hence to be pruned. Hence, in this phase, all the qualified items are labeled with difficulty level under each type.

The second phase is the Intelligent Item assigning phase. The training data set has both pretest score and final class label. When the new test taker comes, based on his score in pretest, the learning ability is predicted using Naïve Bayesian Classification approach. Upon the predicted level, the specific level of item from Type 1 Questions is posed. When there is correct response, next highest difficulty level of item is posed. Then it can be proceed to the next constraint type of Questions. Or else if there is incorrect response, the next lower difficulty level item is posed, proceeded with next lower level of type also.

Thus, the adaptivity in the testing environment can be achieved efficiently which is the critical objective of the Intelligent Tutoring Systems.

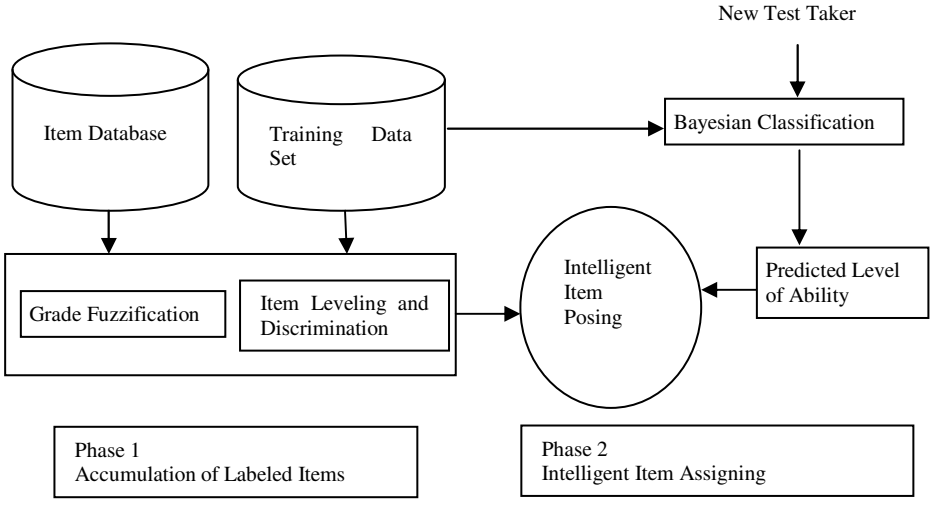


Fig. 3.1. Intelligent Item Posing in Adaptive Testing

Following is the methodology proposed for this study.

Phase 1

Step 1: Classify the Learners in Training Data Set into High, Medium and Low Performers

Step2: Fuzzificate the Grade into High, Medium and Low degree indication for each type of Item individually Step3: Compute Item Difficulty and Item Discrimination using Norm Referencing Approach for each type of Items individually

[Weightage for performance: High=1, Medium=0.5 and Low=0

Let U, M, L be the Number of High, Medium and Low performers.

U_H, U_M and U_L are the sum of the weightage of High, Medium and Low Performance of High Performers.

M_H, M_M and M_L are the sum of the weightage of High, Medium and Low Performance of Medium Performers.

L_H, L_M and L_L are the sum of the weightage of High, Medium and Low Performance of Low Performers.

Let $P_H = (U_H + U_M + U_L)/U$,
$P_M = (M_H + M_M + M_L)/M$ and
$P_L = (L_H + L_M + L_L)/L$]

(3.1) Item Difficulty $IDiff = 1 - [(P_H+P_L)/2]$ (3.2)

Item Discrimination (Medium To Low)

$$IDiscML = P_M - P_L$$

Item Discrimination (High to Medium)

$$IDiscHM = P_H - P_M$$

(3.2) Find all Items with negative Discrimination and prune it.

Phase 2

Step 1: Find the Posterior Probability of H conditioned on X, P(X|H)
[Let X be new Test Taker, H be the Hypothesis that X belongs to the Class C,
P(X) be the Prior Probability of X, P(H) be the Prior
Probability of H,
P(X|H) be Posterior Probability of X conditioned on H]

$$P(X|H) = (P(X|H)*P(H))/P(X)$$

Step 2: Based on the predicted level of ability of the new test taker, propose item from Type1

Step 3: If the response is correct,
If no Items in that Type,
Proceed to next higher Type
Else
Assign Next higher level of Difficulty Item in same Type
Else
If no Items in that Type,
Proceed to next lower Type
Else
Assign next lower level of Difficulty Item
Step 4: Repeat Step3 till the Test concludes.

3.1 Data Gathering

The Data set introduced here consists of the Test taken by 56 students of MCA course during the fifth semester for the subject Data Warehousing and Data Mining. There were 50 questions under all the different types as mentioned above. Questions were delivered via a CBT on Moodle Learning Course Management System. Also, the score in the pretest is also taken which is used to classify and predict the learning ability.

3.2 Accumulation of Labeled Items

The learners in the training data set are classified into 3 classes as high, medium and low performers based on the statistical measure. The numeric grades obtained are transformed into symbolic identification by grade fuzzification. Following is the sample fuzzified data set for first 5 learners for some Items under Type 1 and Type 4.

Table 3.1. Grade Fuzzified Data Set

Item Type	Q Id	Fuzzified Level	Learners 1	2	3	4	5
Type 1	1	Low	0	1	0	0	1
		Medium	0	0	0	0	0
		High	1	0	1	1	0
	15	Low	0	1	1	0	0
		Medium	0	0	0	0	0
		High	1	0	0	1	1
Type 4	14	Low	0	0.5	0	0	0
		Medium	0	0.2	0	0	0.5
		High	1	0	1	1	0.3
	29	Low	0	1	0.5	0.5	1
		Medium	0.5	0	0.2	0.2	0
		High	0.3	0	0	0	0

Item Difficulty is calculated to label the item with difficulty level. Item Discrimination is computed to discard items which are not well discriminating. Here, Norm Referencing approach is used to find those measures. The table 3.2 has the sample item analysis.

Table 3.2. Item Difficulty and Discrimination

Item Type	Q Id	Item Difficulty	Item Level	Item Discrimination		Status of Item
				Medium - Low	High-Medium	
Type 1	2	0.25	Easy	0	0.5	Keep
	42	0	Very Easy	-0.5	0.5	Discard
	50	0.25	Easy	0.25	0.25	Keep
Type 4	14	0.375	Medium	0.5	0.25	Keep
	29	0.813	Very Hard	0.25	0.125	Keep

3.3 Intelligent Item Assigning

When a new test taker comes, based on the pretest score, the class membership is predicted using Naïve Bayesian Classification. The item is posed based on the predicted level of learning ability. Initially, items are posed form Type1 Items. When the learner gives the correct response, the next higher difficulty level of item is posed if there exists, or else item is taken from next higher Type. Similarly, when there is incorrect response, the next lower difficulty level of item is posed if there exists, or else item from next lower type is taken.

Thus, the items are assigned to the learners in an adaptive manner automatically and efficiently. Thereby the main objective of Intelligent Tutoring System is achieved.

4 Conclusion and Further Enhancements

There is a great need in the learning management system area to monitor test results on a large scale as well as to identify questions that are most likely to be benefited by student according to the knowledge level of the student. The applications of item response theory modeling help this issue. Item banking allows for the development of computerized adaptive tests that reduce respondent burden and increases reliable measurement by using a methodology that targets in on a respondent's true score. This study proposes an intensified leveling system using Grade Fuzzification, Norm Referencing Theory and Naïve Bayesian Classification. The level of new learners can be predicted using Bayesian Classification. The items in the pool are leveled on item difficulty using Norm Referencing. In addition to that item discrimination index also used to well differentiate the learners, by considering only those items which very well discriminates the high, medium and low performers. Hence, only efficient and intellectual questions are considered. Therefore, the system poses the intelligent question from the pool based on the predicted level of ability. Since the test is based on their learning ability, test fairness is maintained. Items with different constrained constructs are involved. Hence, effective learning is achieved at the most without any compromise which is the objective of Intelligent Tutoring System. Other type of item analysis model can be tested to give better performance since many techniques are available in IRT with different parameters. The system can be designed in such a way that it can accept other academic attributes to develop the decision tree for classifying well the new learner which can make it a generic one.

References

1. Sokolova, M., Totkov, G.: Accumulative Question Types in Elearning environment. In: International Conference on Computer Systems Technologies – CompSysTech (2007)
2. Lufi, D., Okasha, S., Cohen, A.: Test anxiety and its effect on the personality of students with learning Disabilities. Learning Disability Quarterly 27(3) (2004)
3. Marks, A.M., Cronje, J.C.: Randomised items in computer-based tests: Russia roulette in assessment? Journal of Educational Technology & Society 11(4) (2008)
4. Bennette, R.E.: Construction versus Choice in Cognitive measurement: Issues in constructed response. In: Performance Testing and Portfolio Assessment, pp. 1–27. Lawrence Erlbaum Associates, Hillsdale
5. Erdoğdu, B.: Computer based testing evaluation of question classification for Computer Adaptive testing, A Master Thesis (2009)
6. Lilley, M., Barker, T.: The development and evaluation of a computer-adaptive Testing application for English language. In: Proceedings of the 6th Computer-Assisted Assessment Conference, Loughborough University, United Kingdom (2002)
7. Cheng, S.-C., Huang, Y.-M., Chen, J.-N., Lin, Y.-T.: Automatic Leveling System for E-Learning Examination Pool Using Entropy-Based Decision Tree. In: Lau, R., Li, Q., Cheung, R., Liu, W. (eds.) ICWL 2005. LNCS, vol. 3583, pp. 273–278. Springer, Heidelberg (2005)
8. Izard, J.: Trial Testing and Item Analysis in Test Construction, Module 7 in Quantitative Research Methods in Educational Planning. UNESCO International Institute for Educational Planning (September 2005)

Implementation of Multichannel GPS Receiver Baseband Modules

Kota Solomon Raju[1], Y. Pratap[1,2], Virendra Patel[1,2], Gaurav Kumar[1], S.M.M. Naidu[2], Amit Patwardhan[2], Rabinder Henry[2], and P. Bhanu Prasad[1]

[1] Central Electronics Engineering Research Institute (CEERI)/Council of Scientific and Industrial research (CSIR) Pilani-333031
[2] International Institute of Information Technology Pune
solomon@ceeri.ernet.in,
{pratap.sost.iiit,virendra369,gauravpride1985}@gmail.com,
{mohans,amitp}@isquareit.ac.in, henrysal2000@gmail.com,
bhanu@ceeri.ernet.com

Abstract. Global Positioning Systems are mainly used for finding the location of an object across the globe. GPS receivers can be implemented by using software defined radio techniques. In this paper, hardware implementation of base band (acquisition and tracking) modules of a GPS receiver using system generator 9.2 has been carried out. The implementation will be tested on Lyrtech (small form factor-software defined radio) platform which consists of 3 layers. The upper layer being the radio frequency (ISM band receiver) layer, middle layer is the ADACMasterIII layer and the last is digital processing (DSP) layer. The data transfer between the FPGA Virtex4 SX35 and DSP module is done using a TMS320DM6446 Davinci processor. Generation of 17MHz Intermediate frequency from the RF signal received by the GPS receiver has been achieved. Currently the process of building the base band modules of acquisition and tracking are in progress. Acquisition is being implemented using parallel code phase search algorithm. This is performed by using FFT. Tracking module is implemented by using Costas loop and Delay Lock Loop (DLL). First the base band modules are being made in simulink and simulation results will be tested. Once this is achieved, real time hardware implementation will be done. The results will lead to the development of indigenous GPS receivers with single and multiple channels within the same hardware with reconfiguration. Also it is adaptive for consistent receiving and tracking of the signals.

Keywords: GPS, Base band, Acquisition, Tracking, Software defined radio, Parallel code phase search, FFT, Costas loop, Delay lock loop.

1 Introduction

Global positioning system is a satellite based navigation system which was started in 1973. GPS applications include surveying, space navigation, automatic vehicle monitoring, emergency services dispatching, and mapping and geographic information system geo referencing [1]. At present there are 32 GPS satellites revolving around

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 817–824.
springerlink.com

the globe. Out of these, 24 satellites are currently divided into six orbits and each orbit has four satellites. Each of these orbits makes an inclination angle of 55° with earth's equator. Each of these orbits are separated from each other by 60° thus completing the entire 360°.Each satellite rotates around the earth two times in a sidereal day in their respective orbit having a radius of approximately 26550km [2]. This paper discusses digital base band blocks of GPS receivers and algorithms used in implementation of acquisition and tracking modules. This paper also discusses about the SFF-SDR board which is used for the digital GPS receiver implementation.

2 GPS Signal Structure

A GPS satellite transmits the GPS signal and it is received by the antenna of the GPS receiver. The GPS signals are transmitted on two radio frequencies in the UHF band. The UHF band covers the frequency band from 500MHz to 3 GHz. These frequencies are referred to as L1 and L2 and are derived from a common frequency of 10.23MHz [3]. A GPS signal consists of four main components. These are Coarse Acquisition code, Navigation data, Carrier frequency and Doppler frequency shift.

3 Digital Multichannel GPS Receiver

Once the signal is captured by the antenna of a GPS receiver, through the radio frequency chain the input signal is amplified to proper amplitude and the frequency is converted to desired frequency [2]. Now ADC is used for the digitization of the signal. Once the digital signal is obtained it is down converted to a required intermediate frequency (IF). The below figure shows the multichannel GPS receiver.

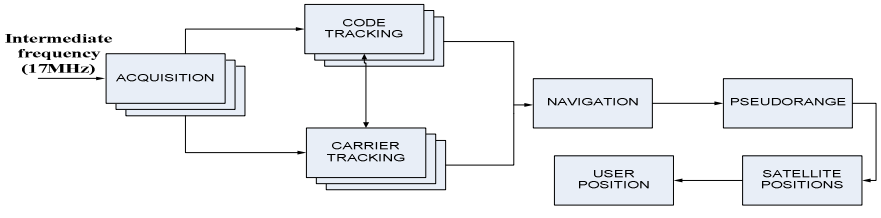


Fig. 1. Multichannel GPS receiver blocks

The IF is then sent to the first digital baseband module i.e. Acquisition module. Acquisition module helps in finding out from which satellite the signal is coming. The tracking module helps in finding out the phase transition of the navigation data [2]. The navigation data gives us information of the orbit of a satellite. From the navigation module we obtain the pseudo-range and the ephemeris data, which gives us the information about the satellite positions.

4 Acquisition Module

The main purpose of acquisition is to determine visible satellites and coarse values of carrier frequency and code phase of the satellite signals.

Algorithms used for implementing acquisition and their comparison [3].

Table 1. Execution time for each of the three implemented acquisition algorithms

Algorithm	Execution Time	Repetitions	Complexity
Serial Search	87ms	41943	Low
Parallel frequency Space search	10ms	1023	Medium
Parallel code Phase search	1ms	41	High

4.1 Parallel Code Phase Search Acquisition Algorithm

The goal of the acquisition is to perform a correlation with the incoming signal and a PRN (pseudo random noise) code. The below figure shows the parallel code phase acquisition algorithm [4].

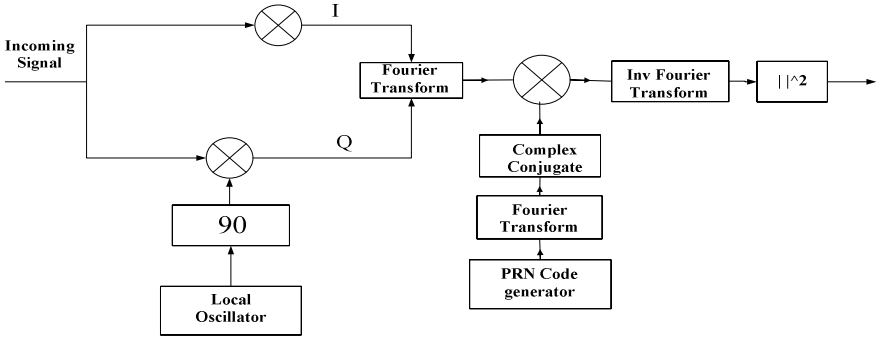


Fig. 2. Parallel code phase search acquisition

Demodulation and dispreading are performed in the parallel code phase acquisition. One provides the demodulation carrier and the other provides the dispreading code [5].

The two main operations performed in the above mentioned algorithm are demodulation and dispreading. Initially a local oscillator generates a local carrier replica and its 90º phase shifted signal. These signals are multiplied by the incoming digitized intermediate frequency signal. This generates the I (in phase) signal and the Q (quadrature) signal respectively .All the energy is stored in the In-phase signal. The I and Q signals are combined to form a complex input signal to the DFT function.

$$x(n) = I(n) + jQ(n) \tag{1}$$

Next comes the dispreading of the signal. The generated PRN code is transformed into the frequency domain and the result is complex conjugated. The Fourier transform of the input is multiplied with the Fourier transform of the PRN code. Actually circular cross correlation is performed. The result of the multiplication is transformed into the time domain by an inverse Fourier transform. The absolute value of the output of the inverse Fourier transform represents the correlation between the input and the PRN code. If the peak is present in the correlation, the index of this peak marks the PRN code phase of the incoming signal [3].

If we take two finite length sequences x (n) and y (n), both with length N, the DFT can be computed as:

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N} \tag{2}$$

$$Y(k) = \sum_{n=0}^{N-1} y(n) e^{-j2\pi kn/N} \tag{3}$$

The circular cross correlation between two finite length sequences x (n) and y (n) both with length N and with periodic repetitions is computed as:

$$z(n) = \frac{1}{N} + \sum_{m=0}^{N-1} x(m) y(m+n) = \frac{1}{N} + \sum_{m=0}^{N-1} x(-m) y(m-n) \tag{4}$$

After omitting the scaling factor 1/N, the discrete N-point Fourier transform of z (n) can be expressed as

$$Z(k) = \sum_{n=0}^{N-1} \sum_{m=0}^{N-1} x(-m)\, y(m-n)\, e^{-j2\pi kn/N} \tag{5}$$

$$\sum_{m=0}^{N-1} x(m)\, e^{-j2\pi kn/N} \sum_{n=0}^{N-1} y(m+n)\, e^{-j2\pi k(m+n)/N} = X^*(k) Y(k) \tag{6}$$

Where

$X(k)$ = Discrete Fourier transform of the finite length sequences x (n)
$Y(k)$ = Discrete Fourier transform of the finite length sequences y (n)
$X^*(k)$=Complex conjugate of X (k)
$Z(k)$ = Discrete N-point Fourier transform of $z(n)$

The code phase and the carrier frequency parameters are further sent to the Tracking module for further refining.

5 Tracking

The main purpose of tracking is to refine the coarse values of the code phase and the frequency and to keep track of these as the signal properties changes over time. It demodulates the incoming signal to obtain the 50Hz navigation data bits.

The tracking mainly consists of two parts.

1. Code tracking (DLL)
2. Carrier tracking(PLL)

The below figure explains the complete tracking module [3].

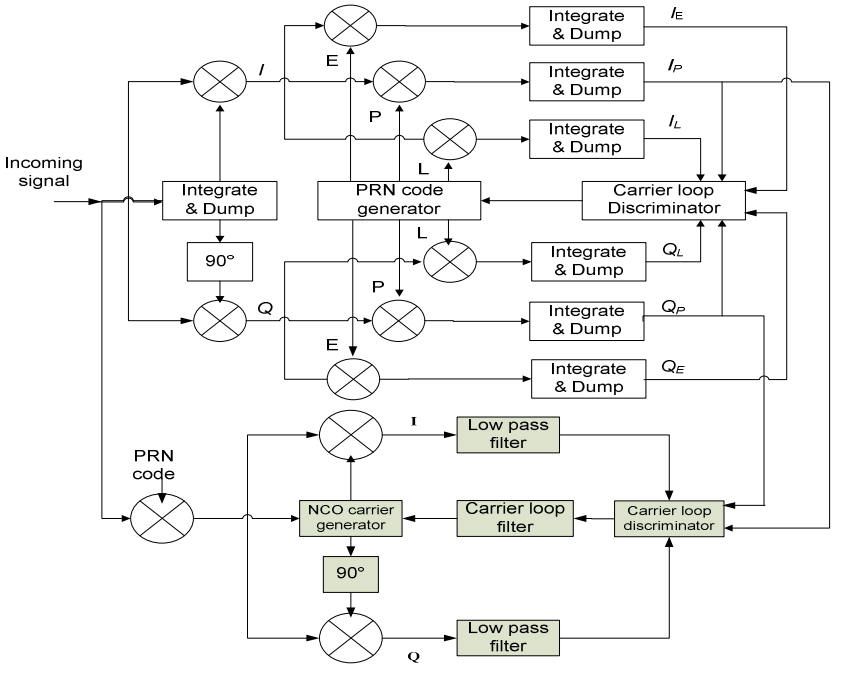


Fig. 1. Block diagram of the combined DLL and PLL tracking loops

Code Tracking

The code tracking loop used in GPS receivers is a delay lock loop (DLL) called an early minus late tracking loop. Here also the DLL discriminator provides the necessary feedback required to ensure the replica signal is always with the incoming signal. The main purpose of the code tracking lop is to keep the track of the code phase of a specific code in the signal. The o/p of the code tracking loop is a perfectly aligned replica of the code. The reason to use DLL is to correlate the I/P signal with three replicas of the code. [Early, Late, and Prompt]. The incoming C/A code is down converted to baseband by multiplying the incoming signal with a perfectly aligned local replica of the carrier wave.

Carrier Tracking

As shown in figure 2, Costas Phase Loop measures the phase error between incoming and local generated carrier, after loop filter the error will be used to adjust frequency of a local oscillator. Costas loop is insensitive to both 180° phase shifts and phase

transitions caused due to navigation bits. The I (in) phase arm of this loop keeps all the energy as given below [3].

$$D^k(n)\cos(\omega_{IF}n)\cos(\omega_{IF}n+\varphi)=\frac{1}{2}D^k(n)\cos(\varphi)+\frac{1}{2}D^k(n)\cos(2\omega_{IF}n+\varphi) \quad (7)$$

φ = phase difference between the phase of the input signal and the phase of the local replica of the carrier phase.

When multiplication is performed in the quadrature arm the above equation changes to:

$$D^k(n)\sin(\omega_{IF}n)\sin(\omega_{IF}n+\varphi)=\frac{1}{2}D^k(n)\sin(\varphi)+\frac{1}{2}D^k(n)\sin(2\omega_{IF}n+\varphi) \quad (8)$$

After low pass filtering the following two signals remain as:

$$I^k=\frac{1}{2}D^k(n)\cos(\varphi) \quad (9)$$

$$Q^k=\frac{1}{2}D^k(n)\sin(\varphi) \quad (10)$$

Phase error of the local carrier phase replica is:

$$\frac{Q^k}{I^k}=\frac{\frac{1}{2}D^k(n)\sin(\varphi)}{\frac{1}{2}D^k(n)\cos(\varphi)}=\tan(\varphi) \quad (11)$$

The phase error is reduced when correlation in the quadrature phase arm is zero having the maximum value in the In-phase arm.

$$\varphi=\tan^{-1}\left(\frac{Q^k}{I^k}\right) \quad (12)$$

6 Implementation of GPS Receiver on Lyrtech SFF-SDR Board

The SFF-SDR board is conceived and designed to be used in the development of the application in the field of software defined radio. The board is composed of three different platforms: Digital Processing Module, ADACMasterIII, RF Module. The digital processing module uses a Virtex-4 SX35FPGA and a TMS320DM6446SoC to implement custom IP and acceleration functions with varying requirements from one protocol to another supported on the same hardware.The ADACMasterIII is equipped with dual channel analog to digital and digital to analog converters.The RF module covers a variety of frequency ranges in transmission and reception. The interface between DSP and FPGA is performed by using TMS320DM6446SoC Davinci processor .The below figure shows the flow of the SFF-SDR board[6].

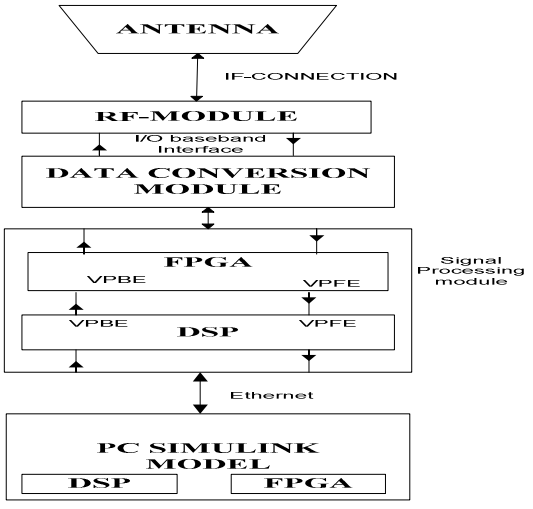


Fig. 2. Design flow of the SFF-SDR board

7 Results and Future Work

We have been able to generate 17MHz Intermediate Frequency required for our board.

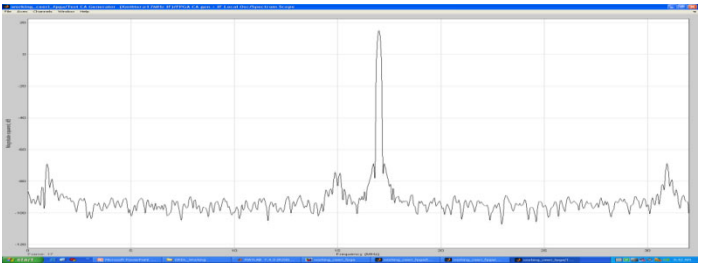


Fig. 3. 17MHz Intermediate Frequency

After generating IF the next task is to implement the Acquisition and Tracking module.

8 Conclusion

This implementation will lead to the development of indigenous GPS receivers with single and multiple channels within the same hardware with reconfiguration.

References

1. Dana, P.H.: Global positioning system (GPS) time dissemination for real time applications, p. 1. University of Texas, Austin (1997)
2. Tsui, J.B.-Y.: Fundamentals of Global Positioning System Receivers, 2nd edn., pp. 2, 31, 33. Wiley-Interscience (2005)
3. Borre, K., Akos, Nicolaj, Rinder, Jensen: A Software Defined GPS and Galileo Receiver, Birkhauser, p. 17, 82, 85, 88 (2007)
4. Li, C., Qian, Y., Lu, M., Feng, Z.: The design and implementation of GPS software simulation platform, p. 4. University of Tsinghua, Beijing (2008)
5. Hamza, G., Zekry, A., Motawie, I.: Implementation of a complete GPS receiver using Simulink. IEEE Circuits and Systems Magazine, 45 (2009)
6. User's guide on Lyrtech's Small Form Factor SDR Evaluation Module/Development Platform (October 2008)

Towards a Practical "State Reconstruction" for Data Quality Methodologies: A Customized List of Dimensions

Reza Vaziri[1] and Mehran Mohsenzadeh[2]

[1] Department of Computer Science, Science Research Branch, Azad University of Iran, Tehran Iran
r.vaziri@iauctb.ac.ir
[2] Department of Computer Science, Science Research Branch, Azad University of Iran, Tehran Iran
mohsenzadeh@srbiau.ac.ir

Abstract. Data quality (DQ) has been defined as "fitness for use" of the data (also called Information Quality). A single aspect of data quality is defined as a "dimension" such as "consistency", "accuracy", "completeness", or "timeliness". In order to assess and improve data quality, "methodologies" have been defined. Data quality methodologies are a set of guidelines and techniques that are designed for assessing, and perhaps, improving data quality in a given application or organization. Most data quality methodologies use a pre-defined list of dimensions to assess the quality of data. This pre-defined list is usually based on previous research and may not be related to the specific application at hand. As a prelude (or state reconstruction phase) for methodologies, a useful list of dimensions specific to the current application or organization must be collected. In this paper we propose a state reconstruction phase in order to achieve that.

Keywords: Data Quality, Dimensions, Methodology.

1 Introduction

Data quality has been defined as "fitness for use" of data. In order to assess and measure data quality in various contexts many data quality methodologies have been developed. A methodology refers to a set of guidelines and techniques that are designed for the assessment and improvement of data quality in a specific application or organization. In [Batini 2009] methodologies have been divided into three main "phases and steps". The three steps and phases are the following:

1. State reconstruction: which collects contextual information on organizational data, processes and services.
2. Assessment / Measurement: which measures the quality of data along relevant "dimensions". The term "measurement" refers to measuring the values of data itself, and the term "assessment" refers to comparison against reference values.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 825–835.
springerlink.com

3. Improvement: which proposes techniques and strategies for reaching higher levels of data quality, perhaps levels specified by the organization's management.

Remember that a dimension is a single aspect of data quality such as "accuracy". Notice, that the "Improvement / Assessment" phase measures the quality of data along "relevant dimensions". However, the essential question that remains is that "What are the relevant dimensions?" What this paper is trying to do is to define a list of dimensions that are specifically designed for the application or organization under data quality study. The basic assumption is that a pre-defined of list of dimensions may not necessarily address the issues and needs of the specific organization that is being studied. The identification of such list of dimensions is part of a new practical state reconstruction. This paper is organized as follows: section 1 is the Introduction. In section 2 is the Related Work about dimensions and methodologies. In section 3 is the proposed solution where a new questionnaire-based method is introduced to define a list of dimensions specific for the application or the organization at hand. Section 4 is the evaluation, and Section 5 is the conclusion and future research.

2 Related Work

In order to assess and measure data quality in various contexts many data quality methodologies have been developed. A methodology refers to the specific set of guidelines and techniques that are designed for the assessment and improvement of data quality in a specific application or organization. Most data quality methodologies that have been proposed are geared toward a specific type of organization or information system. For example, in [Scannapieco 2004] a methodology has been proposed, namely DaQuinCIS, which is specifically designed for the Cooperative Information System (CIS). In [Long and Seko 2005] CIHI (Canadian Institute for Health Information methodology) is introduced which is a methodology specifically designed for Canadian health institutions. There seems to be a lack of methodologies that are general enough to be applied to most data quality situations, regardless of the type of organization, information system, etc. Also these methodologies must be simple enough, so that they could be understood and implemented by the people of various backgrounds, who may not necessarily be data quality or IT experts.

Data quality dimension has been defined as a set of data quality attributes that represents a single aspect or construct of data quality [Wang 1996]. Data quality dimensions have been the subject of much debate in the data quality literature. Various lists of dimensions with different definition and classification have been proposed [Wang 1996, Wand 1996, Jarke 97, Goodhue 95, Delone 92, Ballou 85, Zmud 78 etc]. For instance, In [Wang 1996] some of the more prominent dimensions are defined as the following:

Table 1. Wang and Strong List of Data Quality Dimensions [Wang 1996]

Dimensions	Definition
Accessibility	The extent to which data is available, or easily or quickly retrievable
Appropriate Amount	The extent to which the volume of data is appropriate for the task at hand
Believability	The extent to which data is regarded as true and credible
Completeness	The extent to which data is not missing and is of sufficient breadth and depth of the task at hand
Concise Representation	The extent to which data is compactly represented
Consistent Representation	The extent to which data is presented in the same format
Ease of Manipulation	The extent to which data is easy to manipulate and apply to different tasks
Free-of-Error	The extent to which data is correct and reliable
Interpretability	The extent to which data is I appropriate languages, symbols, and units, and the definitions are clear
Objectivity	The extent to which data is unbiased, unprejudiced, and impartial
Relevancy	The extent to which data is applicable and helpful for the task at hand
Reputation	The extent to which data is highly regarded in terms of its source or content
Security	The extent to which access to data is restricted appropriately to maintain its security
Timeliness	The extent to which the data is sufficiently up-to-date for the task at hand
Understandability	The extent to which data is easily comprehended
Value added	The extent to which data is beneficial and provides advantages from its use

Not only different lines of research differ in the list of the proposed dimensions, they also differ in the definition of the dimensions itself. For instance, [Wang 1996] defines "Accuracy" as the extent to which data are correct, reliable, and certified, however, [Ballou 1985] defines accuracy as the extent to which database values correspond to real world values.

Data quality dimensions are generally divided into four main categories: **1. Intrinsic 2. Contextual 3. Representational 4. Accessibility.**

Intrinsic: These are the qualities that data possesses intrinsically, that is, qualities that are part of fundamental nature of the data. For instance, accuracy, or how close a data value is to its real world value, is usually considered an intrinsic quality.

Contextual: These are qualities that are important considering the "Context" or background in which the data is used. For instance, "Timeliness" for the task of stock management is limited to a few seconds, but for the tasks related to the post-office it could vary from days till weeks.

Representational: These are qualities that relate to how well the data could be represented in a computer system. For instance, "Understandability" and "Conciseness" are such qualities. In other words, can the data be represented in a computer system (such as a database) so that it could be both understandable and concise to the users?

Accessibility: These are qualities that relate to how well the data could be accessed within a computer system. For instance, how fast is the data accessible and how secure is it?

In a very useful study in [Lee 2002] the major classification of quality dimensions from earlier research have been collected and compared against each other. The comparison is shown in Table 2. Among other things this table shows the variety of opinions that exist among researchers for what makes a complete and relevant set of dimensions. As stated in the previous sections, sometimes there is not even a general agreement on the meaning of a particular dimension. Finally, notice researchers sometimes do not agree on the "category" of the dimension either. For instance, [Ballou 1985] and [Wang 1996] put "completeness" as part of the "Contextual" quality of data, whereas, [Jarke 1997] places it in the "Intrinsic" category. "Completeness" as an intrinsic quality could be defined in terms of any missing value at all from a database, but as a "Contextual" quality it could be defined as missing value among those values that are needed or used by the data users.

In [Lee 2002] also the AIMQ methodology has been introduced. In this questionnaire-based methodology for each dimension several "items" or questions are developed. For example, for "completeness" the following items may be used:

Completeness. (6 items)
This information includes all necessary values.
This information is incomplete. (R)
This information is complete.
This information is sufficiently complete for our needs.
This information covers the needs of our tasks.
This information has sufficient breadth and depth for our task.

Based on survey-based answers of the above items the value of a dimension is determined for an organization. The questionnaire-based methodology has the advantage of being both "general" and "simple".

But how are the data quality dimensions identified? For instance, the list introduced by [Wang 1996] is one of the most widely used in data quality literature. The paper selects the dimensions in the following fashion.

Table 2. Data Quality Dimensions proposed by different researchers [Lee 2002]

	Intrinsic	**Contextual**	**Representational**	**Accessibility**
Wang and Strong	Accuracy, believability, reputation, objectivity	Value-added, relevance, completeness, timeliness, appropriate amount	Understandability, interpretability, concise representation, consistent representation	Accessibility, ease of operations, security
Zmud	Accurate, factual	Quantity, reliable/timely	Arrangement, readable, reasonable	
Jarke and Vassiliou	Believability, accuracy, credibility, consistency, completeness	Relevance usage, timeliness, source, currency, data warehouse currency, non-volatility	Interpretability, syntax, version control, semantics, aliases, origin	Accessibility, system availability, transaction, availability, privileges
Delone and McLean	Accuracy, precision, reliability, freedom from bias	Importance, relevance, usefulness, informativeness, content, sufficiency, completeness, currency, timeliness,	Understandability, readability, clarity, format, appearance, conciseness, uniqueness, comparability	Usableness, quantitativeness, convenience of access
Goodhue	Accuracy, reliability	Currency, level of detail	Compatibility, meaning, presentation, lack of confusion	Accessibility, assistance, ease of use (of h/w, s/w), locatability
Ballou and Pazer	Accuracy, consistency	Completeness, timeliness		
Wand and Wang	Correctness, unambiguous	Completeness	meaningfulness	

This list of dimensions has been proposed according to a survey-based study of data quality subject and data consumers. Two surveys are used in the study. The first survey produces a list of possible data quality attributes. These are attributes that come to mind when a typical data consumer thinks of data quality. The second survey assesses the importance of these possible data quality attributes to data consumers. Special factor analysis was done on the assessment from the importance rating survey to produce a set of data quality dimensions that were important to data consumers.

In another set of surveys subjects were asked to categorize the data quality dimensions which after proper analysis and refinements lead to the following four major categories for data quality dimensions: Intrinsic, Contextual, Representational, and Accessibility.

Producing the list of dimensions based on a consumer survey is an innovative idea, but several objections can be raised to the way it was done by [Wang 1996]. The survey was done on subjects who were randomly selected from various experiences. As the paper itself states "subjects should be data consumers who have used data to make decisions in diverse contexts within organizations". The problem is that in different industries, organizations, or even units within an organization the list of dimensions could vary. For a stockbroker the "timeliness" of data is almost everything. In a few seconds the value of stocks could change considerably. However, for a post office the "timeliness" of addresses could vary from days to weeks, since household addresses do not change frequently and not everyone gets important mail every day. Therefore, we propose that in a comprehensive methodology a customized "list of dimensions" for any specific organization be determined. The list that [Wang 1996] proposes has been collected by surveying a random group of data consumers whose needs and experiences may not necessarily match the organization at hand.

The results from the [Wang 1996] research support the above argument as well. The results of the second survey (i.e. importance measurement) show that most of 118 "attributes" (i.e. proposed dimensions) had "a full range of values from 1 to 9, where 1 means extremely important and 9 not important". This could be an indication that the subjects may have different opinion about the importance of each dimension. This as mentioned earlier could be related to the fact that different organizations view dimensions differently in terms of importance.

3 Solution: Creating an Application-Specific List of Dimensions

As mentioned in section 2 using a pre-defined list of dimensions to assess data quality may not address the specific needs of an organization. But the question that remains is that, what is the best way to select the most "relevant dimensions" for an organization? As part of the "state reconstruction phase", we propose a questionnaire to be given to the appropriate subjects who are somehow related to the data processes in the organizations. We propose the following group of subjects in order to obtain different perspectives of the data quality:

Information Professionals (IP's): These are the people who collect and maintain the information for an organization. They are also responsible for designing the systems where information resides.

Information Consumers (IC's): These are the people who use the information.

Independent Experts (IE's): These are defined as experts that have appropriate amount of practical or academic experience in the practices of the organization being evaluated. Also they are called independent because they have no vested interest in the organization being evaluated and thus can present an unbiased opinion.

IE's were included in the surveys for the following reasons. The IP's within an organization might be influenced by the organization policies, years of service, or a

Name: Family Name:

Industry : Organization:

Department: Role: IP, IC, IE:

Job Title: Years of Experience:

How important is each of the following dimensions to this organization?

Accessibility: The extent to which information is available, or easily and quickly retrievable.

Irrelevant Important Vital

0 1 2 3 4 5 6 7 8 9 10

Appropriate Amount: The extent to which volume of information is appropriate for the task at hand.

Irrelevant Important Vital

0 1 2 3 4 5 6 7 8 9 10

Believability: The extent to which information is regarded as true and credible.

Irrelevant Important Vital

0 1 2 3 4 5 6 7 8 9 10

Completeness: The extent to which information is not missing and is of sufficient breadth and depth for the task at hand.

Irrelevant Important Vital

0 1 2 3 4 5 6 7 8 9 10

.

.

.

"Are there any DQ dimensions missing from the above list that could prove useful or important for your organization? Please list them, define them, and rank them from the scale of 0 to 10 just as above."

Fig. 1. Dimension Identification Questionnaire

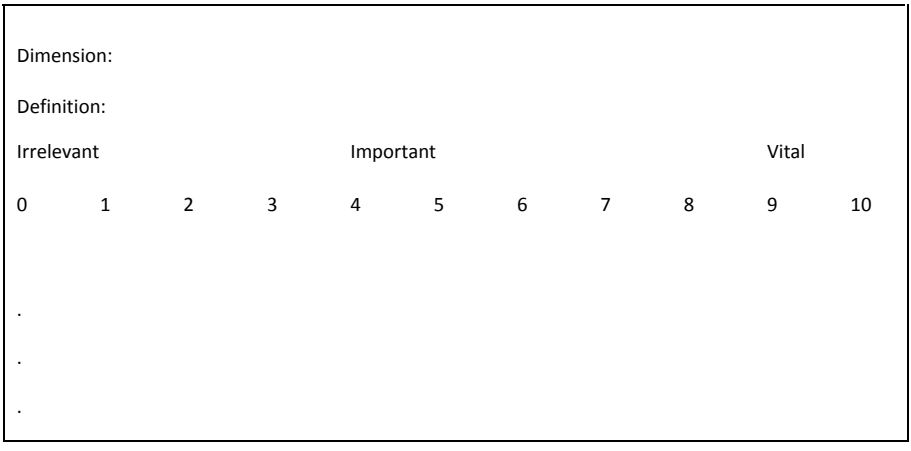


Fig. 1. (*Continued*)

sense of belonging that IP's and IC's might develop towardsan organization. This is not true for IE's and their assessment might identify new DQ problems. Also the IC's and IP's of an organization may not have the proper expertise to assess the quality level of information. IE's, however, due to their long term experience in the practices of the organization must have the proper expertise.

Another objection that could be raised to the [Wang 1996] survey is that the "definition" of the dimensions is not presented in the questionnaire. Some subjects may not know the exact meaning of a definition and also different subjects may interpret them differently as it has been mentioned in the paper itself: "Since we did not include definitions with the attributes, it is possible that data consumers responding to the surveys could interpret the meanings of the attributes differently." Hence, in order to avoid that in our survey we will definitely include the definition of each dimension in the questionnaire.

First, the questionnaire covers all the [Wang 1996] dimensions as a basic set where most organizations need to pay attention to. There is no need to start from scratch again. The basic assumption is that some dimensions are so important and prevalent, that they prove relevant to almost all organizations. The [Wang 1996] list probably covers most of these dimensions. In case a specific dimension is not appropriate for the organization at hand, with a grade of "zero" the subjects could declare it as "Irrelevant". The scale of the questionnaire goes from 0 to 10, where 0 is interpreted as "irrelevant", 5 as "important" and 10 as "Vital". At the end of the questionnaire, the following question is presented to the subjects so that any possible missing dimensions for the specific organization are included.

"Are there any DQ dimensions missing from the above list that could prove useful or important for your organization? Please list them, define them, and rank them from the scale of 0 to 10 just as above."Figure 1 shows a sample questionnaire:

4 Evaluation

A customized list of dimensions could vary considerably for different organizations. Remember the stock market and post office example from section 2. However, other interesting cases could be studied. Consider an auto manufacturing, a dairy production, and a hospital. It's easy to see that the list of dimensions for these organizations could vary significantly. However, there must be dimensions that probably appear high on the list of every organization, for instance "accuracy".

Also it is important to pay attention to the "differences of opinion" between the subjects: IP's, IC's and IE's. For instance, if there is a large difference between what IP's and IC's see as the relevant dimensions, this could mean that the Professional people of the organization are simply not aware of the consumers' needs and concerns. Also if there is a large difference between IE's and the other two subjects (IP's and IC's), this could mean that there is a lack of expertise about the services and products of the organization.

Another practical application that a customized list dimension might have is the "ranking" of the dimensions within an organization. This is especially useful when an organization does not have the resources to deal with all the dimensions at once. Then, it can start from the most important dimension and depending on the available resources work its way down the list.

5 Conclusion

In this paper we introduced the fundamental concepts of data quality, such as dimensions, methodologies and their phases. Then we tried to analyse a common short-coming among the current data quality methodologies. More specifically, methodologies often ignore the fact that a particular application or organization may have its own specific list of dimensions. In other words, a set of dimensions that are useful for one organization may not necessarily be useful for another. Hence, using a pre-defined or pre-selected list of dimensions may not be the best strategy. As part of the state reconstruction phase of the methodologies we proposed a questionnaire-based method to question the appropriate subjects and come up with a relevant list of dimensions. The three groups of subjects are Information Professional, Information Consumers, and Information Experts. The questionnaires were also carefully designed to clearly define each dimension and allow the subjects to rank each dimension or even delete or propose new ones. The useful results from such an approach are the following:

1. Ranking of the dimensions for an organization. Notice that this approach not only provides the list of most relevant dimension, but also it ranks the dimensions according to importance for that organization, something that was rarely done in the previous research. The ranking of the dimensions could prove useful if an organization has limited resources for data quality improvement, and thus, can concentrate on the important dimensions first.

2. Comparison to other organizations: Once the list of most relevant dimensions for various organizations is developed they could be compared against each other. The comparison provides an insight into how different organizations view data quality differently. Also it may identify several dimensions that consistently appear as relevant for most organizations. A list of such dimensions could probably serve as a base or starting point for most data quality methodologies and management processes.

3. Differences of the opinion between groups of subjects: Since we used three distinct groups of subjects (IP', IC's, and IE's) the differences of the opinion between these groups could lead to practical conclusions. For instance, if IE's have a large difference of opinion with IP's or IC's there must be a lack of expertise within the organization and its consumer pool. Also if there is a large difference between IP's and IC's the information professionals in the organization are not aware of their consumers' needs. How such differences of opinion appear in different organizations of various industries may also be interesting. Perhaps in organizations that are involved with a lot of technical work (i.e. IT, Electronics) "expertise" plays an important role, and in non-technical organizations (i.e. dairy products) it will be less important.

In the future research a customized list of dimensions could be prepared for organizations of various industries in order to evaluate the above issues. Also a complete questionnaire-based methodology could be designed to include *measurement/assessment* and perhaps *improvement* phases so that it could be used as a general and simple-to-use data quality methodology.

References

[1] Ballou, D., Pazer, H.: Modeling data and process quality in multi-input, multi-output information systems. Manag. Sci. 31(2) (1985)

[2] Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41(3), Article 16, 52 (2009)

[3] Delone, W.H., McLean, E.R.: Information systems success: the quest for the dependent variable. Information Systems Research 3(1), 60–95 (1992)

[4] Goodhue, D.L.: Understanding user evaluations of information systems. Management Science 41(12), 1827–1844 (1995)

[5] Jarke, M., Vassiliou, Y.: Data warehouse quality: a review of the DWQ project. In: Proceedings of the Conference on Information Quality, Cambridge, MA, pp. 299–313 (1997)

[6] Lee, Y.W., Strong, D.M., Kahn, B.K., Andwang, R.Y.: AIMQ: A methodology for information quality assessment. Inform. Manage. 40(2), 133–460 (2002)

[7] Long, J., Seko, C.: A cyclic-hierarchical method for database data-quality evaluation and improvement. In: Wang, R., Pierce, E., Madnick, S., Fisher, C.W. (eds.) Advances in Management Information Systems-Information Quality Monograph (AMISIQ) Monograph (April 2005)

[8] Scannapieco, M., Virgillito, A., Marchetti, M., Mecella, M., Baldoni, R.: The DaQuinCIS architecture:a platform for exchanging and improving data quality in Cooperative Information Systems. Inform. Syst. 29(7), 551–582 (2004)

[9] Wand, Y., Wang, R.Y.: Anchoring data quality dimensions in ontological foundations. Communications of the ACM 39(11), 86–95 (1996)
[10] Wang, R.: A product perspective on total data quality management. Comm. ACM 41(2) (1998)
[11] Wang, R., Strong, D.: Beyond accuracy: What data quality means to data consumers. J. Manage. Inform. Syst. 12, 4 (1996)
[12] Zmud, R.: Concepts, theories and techniques: an empirical investigation of the dimensionality of the concept of information. Decision Sciences 9(2), 187–195 (1978)

Authors

Reza Vaziri has obtained his B.S. and M.S. in Computer Science from the University ofPittsburgh, Pennsylvania in 1994 and 1996. Currently he is an Assistant Professor at the Azad Universityof Iran (Central Tehran Branch), and also a third year PhD student at the Azad University of Iran(Science and Research Branch). He is writing his Ph.D. thesis on data quality under the supervision of the advising professor Dr. Mehran Mohsenzadeh. r.vaziri@iauctb.ac.ir

Mehran Mohsenzadeh has received his Ph.D. in software engineering from Department of Computer engineering, Science and Research Branch, Islamic Azad University of Iran, in 2004. His major interests are Data Integration Data Quality, Data/Web Mining, Software Architecture, Methodologies, Ontological Engineering, and has published more than 50 conference papers and 5 journal papers. He is Assistant Professor of Computer Engineering Department, Science and Research Branch, Islamic Azad University of Iran. mohsenzadeh@srbiau.ac.ir

A Hybrid Reputation Model through Federation of Peers Having Analogous Function

G. Sreenu and P.M. Dhanya

Department of Computer Science
Rajagiri School of Engineering & Technology, Rajagiri valley, Cochin, India
gsreenug@gmail.com,
dhanya.rajeshks@gmail.com

Abstract. The widespread usage of peer to peer (P2P) system is visible in all major application areas like, file sharing , high performance computing, P2P TV, P2P IP TV etc. Conversely, the foremost confront in the growth of P2P system is the dearth of an efficient scheme to handle malicious and mysterious nodes. The QoS factor of the system will be deeply reduced by freeriders. To facilitate the augmentation, the system should be able to trim down the influence of malicious nodes with a feasible, proficient and scalable reputation model. Trust value computation among peers outline the basis of reputation model. This paper recommends a hybrid reputation model through federation of peers having analogous function. The grouping of peers having analogous function is introduced to reduce the unnecessary searching. The competence of the model can be augmented by the application of following methods. Application of election algorithm with suggested modification is a solution to central point failure. The ring structure formation of peers having analogues function reduces the search time through the network. In addition to that the behavior prediction of peers is made more accurate using the Markov chain.

Keywords: Reputation, Election algorithm, Markov Chain.

1 Introduction

The conventional internet has becoming increasingly occupied with its centralized server architecture. This crowded nature of internet leads to the belief that federal nature should be replaced by decentralized architecture. Shared usage of massive data can be effectively done by P2P networks.P2P networks are becoming increasingly popularized with its scalable, open and anonymous nature. The anonymous nature of P2P network without a centralized influence has paved the way to different intruders like viruses and malwares. The continuous interactions like data sharing and disk space utilization with strange peers consider security as an unavoidable concern. The paper discusses the main solutions existing in security harms and finally reaches a hybrid architecture which provides solutions to different security issues related to P2P area.

In a P2P network each peer is capable of managing both client side and server side effort. The absence of hard core rules and regulations makes the system more flexible

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 837–846.
springerlink.com

to attacks. The different types of security attacks [1,2] include eavesdropping, communication jamming, unauthorized access, man in the middle attack, freeriders, denial of service, file corruption, and Sybil attack. The interactions with peers having anonymous nature make the environment more vulnerable. Reputation systems can easily distinguish good peers from bad peers. The reputation systems following the recommendations received from other peers which have past interactions with the corresponding peer. The efficiency of a good reputation system depends on [3] the creation of recommendation, choice of recommenders and finally the examination of different views gathered from other peers. Reputation system can be categorized based on the architecture for which it is developed. The architecture of P2P network can be categorized into centralized structured, decentralized structured and decentralized unstructured. Centralized P2P architecture provides a central directory server which provides the location information about the data present in the P2P network. Structured but decentralized architecture allocate data among peers based on particular hash functions. Pure P2P means completely unstructured without a central coordinator. Pure P2P systems are free from all types of location restrictions. Data can be stored at any location.

The proposed method suggests a reputation model which uses a hybrid topology. The method is intended to enhance the security of P2P system by considering the speed of processing of information. The peer behavior prediction is done by using Markov Chain method.

2 Related Work

P2P security issues have engrossed more research effort in reputation management through trust calculation. The existing approaches for reputation management are listed below. Xrep [4] Consists of a voting mechanism. The peer will enquire the entire network about the reliability of a particular resource peer A. Peers will vote positively or negatively based on its past interaction with peer A. After collecting the entire feedback, the peer behavior is decided and based on the decision the data can be downloaded or not. TrustMe [5] An anonymous protocol for trust management. It uses public key cryptography methods. Reputation value of each peer is stored in Trust-Holding Agent (THA) peer. The THA peer is randomly chosen by the bootstrap server. All other peers are unknown about the THA peer. The querying peer will broadcast its request for reputation value of peer A and the corresponding THA peer will provide the value. In EigenTrust [6] the concept of global trust value is applied. Initially normalized local trust value of peer A is computed by each peer and finally aggregate the local trust values to a global value. Now each peer is having a global trust value of peer A. Dual EigenRep [7] consider the self recommendation behavior of each peer. Considering two reputation values like recommended reputation value and recommending reputation value. Finally the two values will develop different trust communities. Three dimensional based trust management scheme [8] suggests a 3D normalization to show the highly accurate peer performance. Conventional ratio based calculation is replaced with the idea of closeness and the new computation is more flexible.

The different existing methods have its own disadvantages. Most of the solutions are not feasible for a large real network. Voting mechanism present in Xrep and query broadcast in Trustme will create a lot of message transfers across the network. Presence of large number of nodes in real world should consider the speed of each transaction. The proposed method can assure the speed of each transaction within a minimum delay by associating nodes having similar functions and requirements in a particular group.

3 Comparison of Different Existing Methods

Table 1. Comparison of existing methods

Existing Approaches	Mechanism Used	Complexity	Features
Xrep	Voting	Medium	Implementation is easy but require lot of messages
TrustMe	Public key cryptography methods	Medium	Anonymous and require lot of messages
EigenTrust	Global trust value	High	Globally accepted trust value
Dual EigenRep	Self recommendation behavior	High	Highly accurate
Three dimensional based trust management scheme	3D normalization	High	Highly accurate peer performance

4 Proposed Solution

The proposed model suggests a hybrid topology. The model includes a central server which acts as the coordinator as well as an entry point to the network. The topology itself takes the advantage of reduction in search time. The method can be divided into

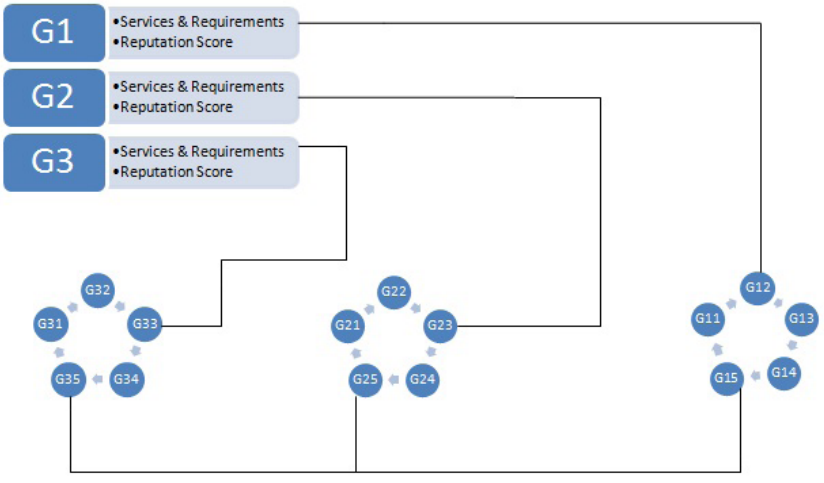


Fig. 1. Architecture of the proposed system

different subsections like access to the system, communication and behavior prediction. The paper describes the method in detailed manner.

4.1 Access to the Network

A server machine is placed as an entry point to the network. The primary function of server is to form groups. These groups are based on Barter System .As per Barter system which was there in our ancient civilization; each person can take a service from other person in exchange of a service from that person. The peers which are satisfying each other's needs and services are classified into one group.

4.1.1 Separate Chaining

The peer classification into different groups is done with the help of separate chaining method. As per this method the server is implemented with the help of a data structure known as array of linked lists. Each location in the array is reserved for each incoming nodes. The corresponding linked list will be used to store the members similar to this peer. The peers which are already part of a group will point to NULL.

The incoming peers will be stored in the array. These peers are then automatically grouped into different groups .The grouping is based on the features of incoming peers. The services provided by a peer and the services needed by a peer form the features of a peer. The peer having similar features is grouped into one group.

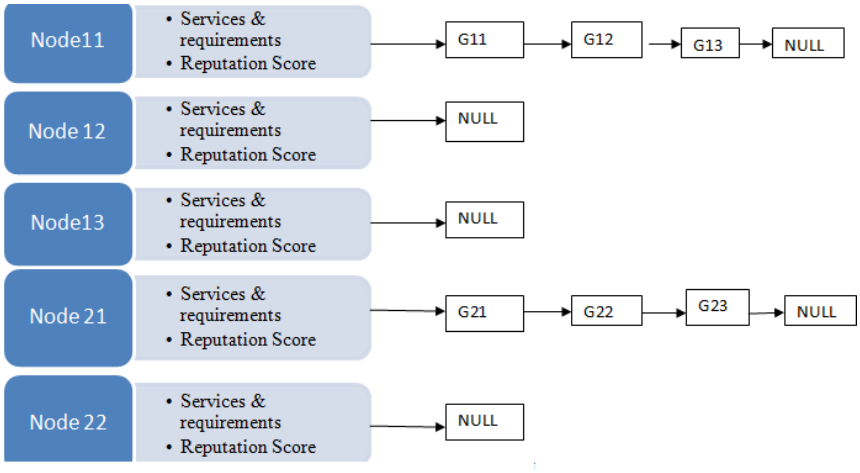


Fig. 2. Node representations

4.1.2 Similarity Score

The similarity score is used to find the similar peers. The features of different peers will be compared and the peers which are having the similarity score above an already set threshold will go into same group. If the similarity score is not satisfying with any of the existing group that peer will form a new group.

4.2 Communication within the Group

The communication is done within the group as well as outside the group. Since the federation is based on similarity of features, most of the requests can be satisfied within the group. The security of information can be ensured within the group by using reputation methods. Each peer will maintain table of information about all peer's trust values. The information can be stored based on the feedback received from peers which are involved in the transaction. The peer having highest trust value will automatically be assigned with a token. The peer which is having a token will act as the coordinator of the group. As the trust value is changing per time the token also passed among the group. The coordinator peer is having the permission to communicate with a member outside the group. Fig 3 will represent a peer group where each peer is having a table of information about trust values of all peers.

4.2.1 Feedback Collection

The feedback should be in the form acknowledgement. If it is a positive acknowledgement the table value corresponding to the peer will be incremented, if it is a negative acknowledgement the value will be decremented. The value stored in the table for each peer will be exchanged and aggregated based on GOSSIP TRUST [9] algorithm. This algorithm is used to maintain a globally accepted value for each peer. If any of the peers having trust value less than the threshold the coordinator will take necessary actions.

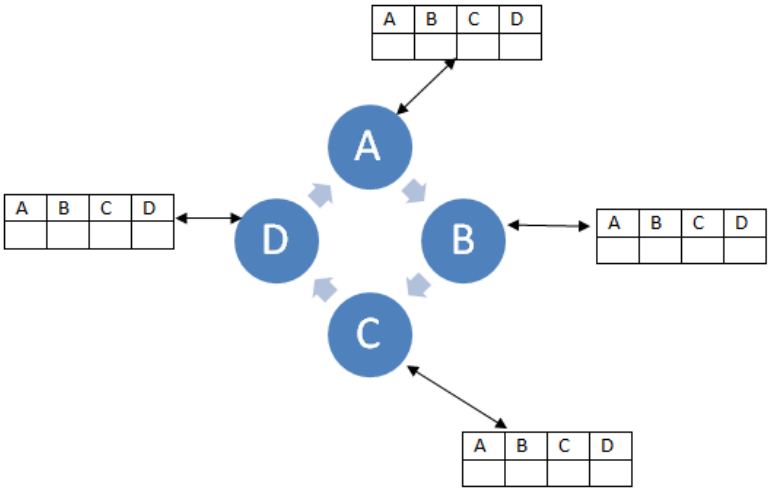


Fig. 3. Group of peers allocated with table of trust values

4.2.2 Blacklisting

The peers with very low trust value will be blacklisted by the coordinator and broadcast this information within the group .So other nodes will reject the peer from communication.

4.3 Communication Outside the Group

If the requests are not satisfied within the group then the communication is between groups. The request will send to each peer coordinator. The peer coordinators will send a reply back to the requestor only if the requested information is present in the group.

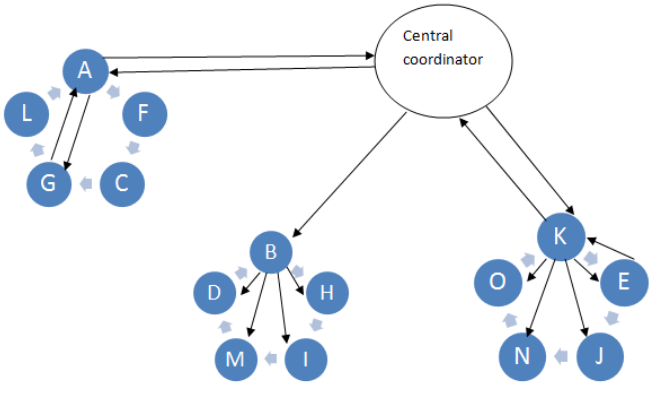


Fig. 4. Communication outside the group

At the time of outside request the coordinator of each group will broadcast the request inside the group. If any of the members within the group gives a reply the coordinator will check the trust value of that peer. If it is a blacklisted one the coordinator will send a negative acknowledgement. Otherwise coordinator will send the address and the trust value of the peer to the central coordinator. The central coordinator will again send back the information to requestor coordinator. If two or more reply came, the peer will choose the one with high trust value. After the transaction the peer trust value will be updated positively or negatively based on the transaction.

4.4 Election of Coordinator

Election of central coordinator or entry point is done at regular interval of time. This election is done to choose a coordinator which act as the backup for central coordinator. The selection is done on the basis that each time the central server will send a request to get the average trust value of each peer group. Then the peer group with high trust value will be allocated to share the information present in the central server. This peer group may be changed at regular intervals based on the trust value. So if anything happens to the central server this peer group coordinator will collect the data from its members and allocated as the next coordinator.

The peer with highest trust value can be identified using Election algorithm [10]. If anything happen to the coordinator present in the peer groups, the peer with next highest trust value will be allocated as the next peer coordinator.

4.5 Behavior Prediction

After collecting the feedback, the behavior of each peer can be identified. Then by applying Markov chain method [11], it is simple to predict the behavior of the peer that whether it will act correctly or maliciously. The Markov property suggests that the future of a state depends on the recent past.

Markov property can be defined like

For any k, $j_0, j_{1,} \ldots, j_{l-1} \varepsilon K$ and any $l >= 1$ and $p >= 0$

$$P(X_{l+p} = s \mid X_0 = j_0, \ldots, X_{l-1} = j_{l-1}) = P(X_{l+p} = s \mid X_{l-1} = j_{l-1})$$

The value of a process at any time in future depends on the most recent past, not on the most primitive past.

4.6 Merits of the System

One of the main advantages of the proposed system is reduction in search time. Since similar peers are in same group unnecessary flooding can be avoided. The application of Markov Chain avoids the storage of past history. This leads to save storage space in trust value calculation.

5 Result Analysis

The proposed model mainly aims to provide security which helps to achieve high query success rate and that in turn increases the peer satisfaction. Here the peer satisfaction not only depends on the successful transaction but also on the time with which the request is satisfied. The factors contributing to the peer satisfaction can be explained with the following chart.

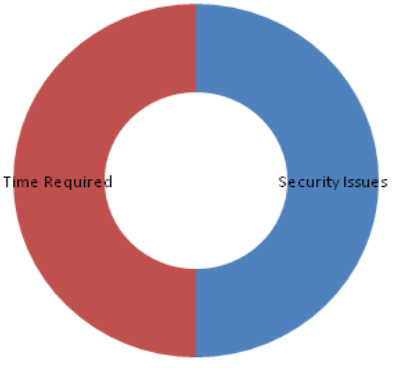


Fig. 5. Factors Contributing to Peer satisfaction

Different security issues should be solved in a timely manner to achieve a high success rate in peer satisfaction. The proposed method can be simulated using a standard simulator. The expected attainment in peer satisfaction can be plotted as shown in the following chart.

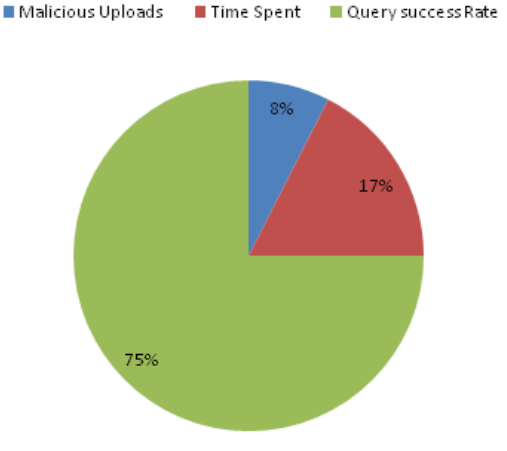


Fig. 6. Expected Result

The above chart describes the expected outcome of the proposed model. The chart shows the percentage of malicious upload, the total time spent and also the query success rate. The small percentage of malicious uploads happens in the initial stage and gradually those peers will be avoided from the system. The time shows the time spent for unnecessary flooding. Small amount of flooding will always occur during inter group communications.

6 Conclusion

The proposed hybrid topology includes the adaptation of many techniques from different areas .The main principle of this method is the grouping of peers based on the similarity of features. The security of data in these groups is ensured by trust value calculation. The trust values are based on the feedback collected from peers in transaction. The chance of central coordinator failure and group coordinator peer is also handled. The implementation of election algorithm is better utilized to handle the election of central coordinator. The prediction method is made simple by the implementation of Markov Chain method. Since the peers are grouped based on the similarity of functions the time spent for entire system processing can be reduced greatly. The main advantage that is expected from this system is reduction in transaction time. Also the division of groups will help to prevent malicious information from moving one group to another.

References

1. Wattenhofer, R.: Attacks on Peer-to-Peer Networks. Dept. of Computer Science, Swiss Federal Institute of Technology (ETH) Zurich (Autumn 2005)
2. Tatsuaki, H., Masahiro, F., Yu, W.: ITU-T Recommendations on Peer-to-Peer (P2P) Network Security
3. Ruohomaa, S., Kutvonen, L., Koutrouli, E.: Reputation Management Survey. In: Second International Conference on Availability, Reliability and Security, ARES 2007 (2007) 0-7695-2775-2/07
4. Damiani, E., Vimercati, S., Paraboschi, S., Samarati, P., Violante, F.: A Reputation-based Approach for Choosing Reliable Resources in Peer-to-Peer Networks. In: ACM Symposium on Computer Communication Security, pp. 207–216 (2002)
5. Singh, A., Liu, L.: Trust Me: Anonymous Management of Trust Relationships in Decentralized P2P Systems. In: IEEE IntI. Conf. on Peer-to-Peer Computing, pp. 142–149 (2003)
6. Kamvar, S., Schlosser, M., Garcia-Molina, H.: The Eigen- Trust algorithm for reputation management in P2P networks. In: Proceedings of the Twelwth International World-Wide Web Conference (WWW 2003), pp. 446–458 (2003)
7. Fan, X., Li, M., Ren, Y., Ma, J.: Dual-EigenRep: A Reputationbased Trust Model for P2P File-Sharing Networks. In: Symposia and Workshops on Ubiquitous, Autonomic and Trusted Computing, uic-atc, pp. 358–363 (2010)

8. Cai, L., Rojas-Cessa, R.: Three-Dimensional Based Trust Management Scheme for Virus Control in P2P Networks. In: Proc. IEEE ICC 2010, CapeTown, South Africa, May 23-27, p. 5 (2010)
9. Zhou, R., Hwang, K.: Gossip-based reputation aggregation for unstructured peer-to-peer networks. In: Proceedings of IEEE International Conference Parallel and Distributed Processing Symposium, pp. 1–10, V3–V23 (2007)
10. Tanenbaum, A.S.: Distributed Operating Systems
11. Norris, J.R.: Markov Chains. Cambridge Series in Statistical and Probabilistic Mathematics

An Amalgam Approach for Feature Extraction and Classification of Leaves Using Support Vector Machine

N. Valliammal and S.N. Geethalakshmi

Assistant Profesor and Associate Professor, Department Computer Science, Avinashilingam Institute of Home Science and Higher Education for Women,
Coimbatore-47, Tamil Nadu, India
Valli.p.2008@gmail.com and sngeethalakshmi@yaho.com

Abstract. This paper describes the need for the development of automatic plant recognition system for classification of plant leaves. In this paper, an automatic Computer Aided Plant Leaf Recognition (CAP-LR) is presented. To implement the above system initially the input image is pre-processed in order to remove the background noise and to enhance the leaf image. As a second stage the system efficiently extracts the different feature vectors of the leaves and gives it as input to the Support Vector Machine (SVM) for classification into plant leaves or tree leaves. Geometric, texture and color features are extracted for classification. The method is validated by K-Map which calculates the accuracy, sensitivity and efficiency. The experimental result shows that the system has faster processing speed and higher recognition rate.

Keywords: Feature Extraction, Classification, Plant recognition, Geometric, Color, Texture features, SVM.

1 Introduction

Botanists need a computer-aided tool without human interaction to study and identify leaves instead of holding a plant encyclopedia. This system guides botanist so that they can quickly search the entire collections of plant Species. Tools are needed to make the botanical information available from the world's herbaria accessible to anyone with a laptop or cell phone. Recently the required data to develop such system is made available. Just by feeding into the computer the photograph of a leaf specimen, the system returns within seconds the top matching species, along with supporting data such as textual descriptions and high resolution type of specimen images. It is also significantly important for environmental protection. The traditional method is time consuming, less efficient and troublesome task. By using our system CAP-LR, a botanist in this field can quickly search the entire collections of plant species within seconds that previously took hours. However, due to the rapid development in computer technologies nowadays, there are new opportunities to improve the ability of plant species identification such as designing a convenient and automatic recognition system of plants.

This work leads towards the impact on the study of biodiversity by identifying plants, primarily using leaf shape. The key issue of leaf recognition is to make sure that the extracted features are stable and can distinguish individual leaves [2]. The

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 847–855.
springerlink.com

identification of different plants species is based on leaf features. The features considered are texture extraction Gray Level Co-occurrence Matrices (GLCM) [7], color and geometric features. However these methods are based on extraction of grayscale images, to use a combination of gray scale and binary texture features. The experimental result proves the effectiveness and superiority of this method.

The paper is organized as follows; Section 2 explores the overview of the system. Section 3 describes the concept of feature extraction. Section 4 explains the working of support vector machine. Section 5 deals with the experimentation and their performance measurement. Finally, the conclusion, future work and references are discussed in section 6.

2 System Overview

This paper overviews the of automatic plant recognition by use of computer which analysis the limitations that exist in the present study and makes several contributions looking foreword to the technology of automatic plant recognition. The work focuses on Feature extraction and classification for plant and tree leaves. The following figure 1 describes the overall structure.

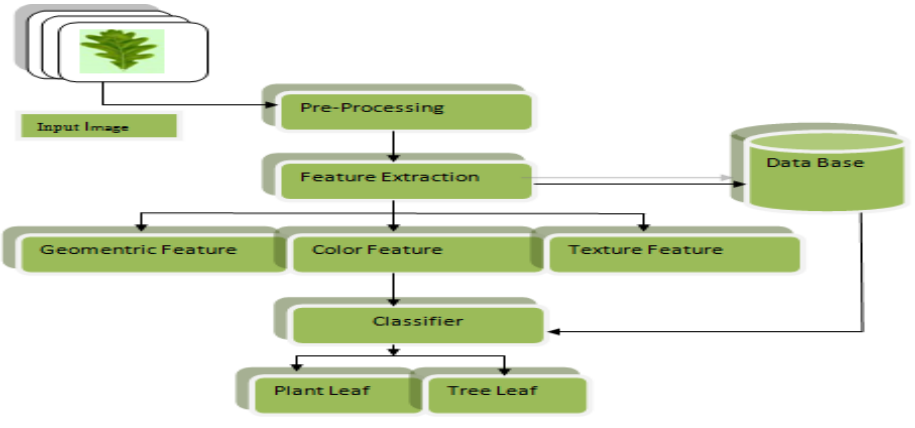


Fig. 1. Overview of the System for classification of plant and tree leaves

The system consists of the following steps, the captured leaf image is given as input. Pre-processing is applied to smoothen and enhance the image to obtain better quality. Features are extracted based on the combination of three divisions, such as geometric, texture and color features after feature extraction the classification process is applied. Finally classification is done using SVM.

3 Feature Extraction

The main work of a leaf recognition system is to extract common features between the images belong with the same class in the images of the data set and consequently

indexing them. The features should carry enough information about the image and should not require any domain-specific knowledge for their extraction. They should be easy to compute in order for the approach to be feasible for a large image collection and rapid retrieval. They should relate well with the human perceptual characteristics since users will finally determine the suitability of the retrieved images.

Feature extraction methodologies analyze leaf images to extract the most prominent features that represent the various classes of objects. The obtained features data are used as inputs to classifiers that assign them to the class that they represent. In this research work the most frequently used Geometric features, GLCM features and Color features are considered to determine the best feature set for leaf database using Region of Interest (ROI) [7], which are the most significant features for classification.

3.1 Geometric Feature

Geometric features are extracted based on the Diameter, Leaf Area, Perimeter and shape of the plant from the plant's outer contour, to describe the overall plant shape. The following basic geometric features are obtained [11], and the values are shown in table 1.

1). Diameter: The diameter is defined as the longest distance between any two points on the margin of the leaf.

2) Leaf Area: The value of leaf area is easy to evaluate, just counting the number of pixels of binary value on smoothed leaf image. The following table shows the results for different geometric features.

3) Leaf Perimeter: Denoted as P, leaf perimeter is calculated by counting the number of pixels consisting leaf margin.

4) Shape Features: Slimness (sometime called as aspect ratio) is defined as follows Slimness = l_1/l_2 , where l_1 is the width of a leaf and l_2 is the length of a leaf.

Table 1. Geometric feature Parameters

Feature	Parameter
Shape	$S = l_1/l_2$
Leaf Area	$LA = \iint I(x, y)dydx$
Leaf Perimeter	$LP = \int \sqrt{x^2(t) + y^2(t)}\,dt$
Diameter	$D = d_2 - d_1$

Table 2. Results obtained from geometric features

Geometric Features	Plant leaf	Tree Leaf
Diameter	213.82	193.964
Leaf Area	23166	10950
Leaf Perimeter	1388.035	885.193
Shape	1.6795	1.0858

3.2 Color Feature

The use of color in plant retrieval is more complicated compared with other applications, since most plants have green tones as their main color. We currently use some basic color features consisting of color histograms and color co-occurrence matrices obtained from the pre-processed image, to represent the color information. Probably the most important aspect for any object is its shape, and the same applies to plants as well. The following table 3 shows the feature parameter for color features and their corresponding feature value is shown in table 4.

Color moments are very effective for Color based image analysis which represents color features to characterize a color image. The information of Color distribution in an image can be captured by the low order moments. The first and second order moment has been proved to be efficient and effective in representing Color distribution of image. Features that are considered normally are mean, standard deviation, skewness, and kurtosis.

Table 3. Color feature parameter

Feature	Parameter
Mean	$\mu \frac{1}{M N} \sum_{i=1}^{M} \sum_{j=1}^{N} P_{ij}$
STD	$\sigma = \sqrt{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} \left(P_{ij} - \mu \right)^2}$
Skewness	$\theta = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} \left(P_{ij} - \mu \right)^3}{M N \sigma^3}$
Kurtosis	$\gamma = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} \left(P_{ij} - \mu \right)^4}{M N \sigma^4}$

Table 4. Results obtained from Color features

Color Features	Plant leaf	Tree Leaf
Mean	1.48+00	1.17E+00
std	2.50E-01	1.39E-01
Skewness	6.20E-02	1.78E+00
Kurtosis	1.00E+00	4.19E+00

3.3 Texture Feature

Besides color and shape, the third core characteristic of an object is its texture. The texture of a plant and tree, formed by the color and vein patterns, is also important in plant identification. In our work, texture features are extracted based on sixteen matrices. The GLCM is a common technique in statistical image analysis that is used to estimate image properties related to second-order statistics.

In this work, the following GLCM features were extracted in our research work: Autocorrelation, Contrast, Correlation, Cluster Prominence, Cluster Shade, Dissimilarity, Energy, Entropy, Homogeneity, Maximum probability , Sum of squares, Sum

average, Sum variance, Sum entropy, Difference variance and Difference entropy. The value obtained for the above features for a typical plant and tree leaf is given in the following table 5.

Table 5. Texture feature parameter

Feature	Parameter
Autocorrelation	$R(s,t) = \frac{E[(x_t - \mu_t)(x_s - \mu_s)}{\sigma_t \sigma_s}$
Contrast	$Contrast = \sum_{i,j=0}^{N-1} P_{ij}(i-j)^2$
Correlation	$correlation = \sum_{i,j=0}^{N-1} P_{ij} \frac{(i-\mu)(j-\mu)}{\sigma^2}$
Cluster Prominence	$\sum_{i=0}^{G-1}\sum_{j=o}^{G-1}\{i+j-\mu_x-\mu_y\}^4 * P(i,j)$
Cluster Shade	$\sum_{i=0}^{G-1}\sum_{j=o}^{G-1}\{i+j-\mu_x-\mu_y\}^3 * P(i,j)$
Dissimilarity	$D = \sum_i \sum_j \lvert i-j \rvert P(i,j)$
Energy	$Energy = \sum_{i,j=0}^{N-1} (P_{ij})^2$
Entropy	$Entropy = \sum_{i,j=0}^{N-1} (P_{ij}) P_{ij}$
Homogeneity	$Homogeneity = \sum_{i,j=0}^{N-1} \frac{P_{ij}}{1+(i-j)^2}$
Angular Second Moment	$ASM = \sum_i \sum_j P(i,j)^2$
Sum of squares	$SOS = \sum_i \sum_j (i-\mu)^2 P(i,j)$
Sum average	$SA = \sum_{i=2}^{2N_g} i g_{x+y}(i)$
Sum variance	$SV = \sum_{i=2}^{2N_g} (i-sa)^2 g_{x+y}(i)$
Sum entropy	$SE = \sum_{i=2}^{2N_g} i g_{x+y}(i) \log\{g_{x+y}(i)\}$
Difference variance	DV=Variance of g_{x-y}
Difference entropy	$DE = \sum_{i=0}^{N_g-1} i g_{x-y}(i) \log\{g_{x-y}(i)\}$

Table 6. Results obtained from Texture features

Texture	Plant leaf	Tree Leaf
Autocorrelation	2.19E+00	1.96E+00
Contrast	1.82E-02	1.48E-02
Correlation	9.58E-01	9.63E-01
Cluster Prominence	1.08E+00	1.14E+00
Cluster Shade	2.95E-01	1.03E+00
Dissimilarity	1.82E-02	1.48E-02
Energy	5.30E-01	2.90E+00
Entropy	7.32E-01	6.80E-01
Homogeneity	4.55E+00	9.93E-01
Angular Second Moment	6.08E-01	6.71E-01
Sum of squares	1.56E+00	1.37E+00
Sum average	2.80E+00	2.64E+00
Sum variance	5.24E+00	4.72E+00
Sum entropy	2.06E+00	6.70E-01
Difference variance	1.82E-02	1.48E-02
Difference entropy	3.47E+00	7.68E-02
Autocorrelation	2.19E+00	1.96E+00
Contrast	1.82E-02	1.48E-02
Correlation	9.58E-01	9.63E-01
Cluster Prominence	1.08E+00	1.14E+00
Cluster Shade	2.95E-01	1.03E+00
Dissimilarity	1.82E-02	1.48E-02

4 Support Vector Machine

This section introduces some basic concepts of SVM and the selection of parameters using SVM. SVMs are set of related supervised learning methods used for classification and regression. SVM is based on statistical learning theory developed to solve pattern recognition problems [6], [8]. It is a supervised classification widely used in different fields. They belong with a family of generalized linear classification. A special property of SVM is, it simultaneously minimize the empirical classification error and maximize the geometric margin. So it is also called Maximum Margin Classifiers. The goal of SVM is to classify two categories as clearly as possible.. For implementing SVM on image classification, a certain number p of training data is given where each data has two parts: the n-dimensional vector of image features and the corresponding labels of data (either 1 or - 1). Each x_i is a n-dimensional vector. The calculation is carried out as follows,

$$S = \left\{ (x, y) \mid x_i \in R^n, y_i \in \{-1,1\} \right\} p_{i=1}$$

5 Experimental Results

In this work, the geometric feature, texture features and color features are extracted from different plant and tree leaves. These combined features are trained with K-Nearest Neighbours (KNN) and SVM classifier. Large image data's are taken and 26 features are extracted for classification. 70% of images are used for training and 30% of images are used for testing purpose. Some sample dataset considered for classification is shown in the following figure 2.

The effectiveness of the system has been estimated using the following measures: Accuracy= (TP+TN)/ (TP+TN+FP+FN), Sensitivity= TP/ (TP+FN) and Specificity= TN/ (TN+FP), where True Positive (TP), True Negative (TN), False Negative (FN) and False Positive (FP) are the total number of Positive choices, the negative choices are the total number of False Positives and True Negatives respectively. Accuracy is the proportion of correctly identified images from the total number of images. Sensitivity measures the ability to identify anomalous images. Specificity measures the ability of the method to identify normal images [22, 23].

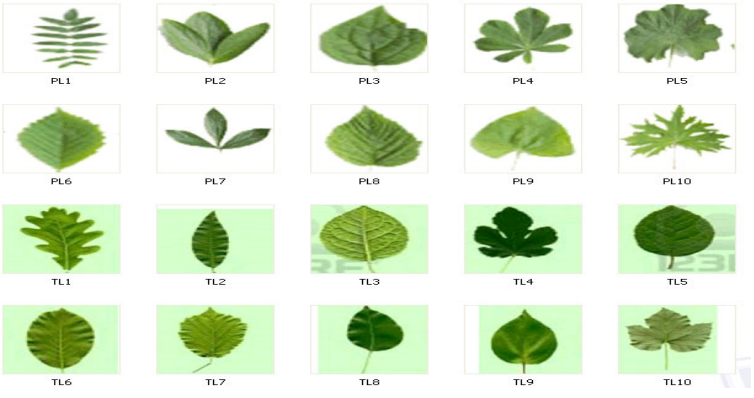


Fig. 2. Sample Plant Leaf and Tree Leaf Dataset

Table 7. Parametric results for classification

S.No	Classifiers	Accuracy	Sensitivity	Specificity
1.	KNN	80%	75%	82%
2.	SVM	85%	78%	86%

The above figure 3 and table 7 shows the accuracy,sensitivity and specificity of KNN and SVM classifier performance. The SVM method produces efficient and suitable results compared to the KNN classifier.

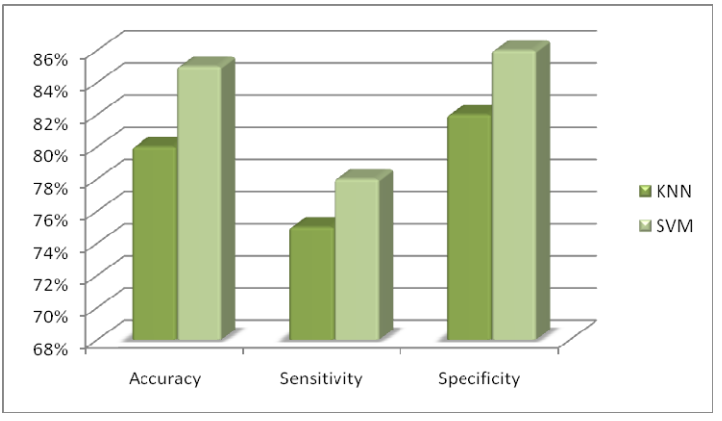


Fig. 3. Results obtained using Accuracy,Sensitivity and Specificity

6 Conclusion and Future Scope

This paper proposes a system for plant and tree leaf identification using geometric, texture and color features of their leaves. The geometric, texture and color features are extracted and SVM is used for classification. This system efficiently extracts the feature vectors of the leaves and classifies into plant and tree leaves by applying the SVM. The method is validated by K-Map which calculates the accuracy, sensitivity and efficiency. The experimental results show that the method has higher recognition rate and faster processing speed.

Future work involves research along the following directions: (1) combining more geometric features and (2) feature selection and reduction method will be done for improving recognition accuracies and classification time.

References

1. Singh, K., Gupta, I., Gupta, S.: SVM-BDT PNN and Fourier Moment Technique for Classification of Leaf Shape. International Journal of Signal Processing, Image Processing and Pattern Recognition 3(4), 68–78 (2010)
2. Chaki, J., Parekh, R.: Plant Leaf Recognition using Shape based Features and Neural Network classifiers. International Journal of Advanced Computer Science and Applications (IJACSA) 2(10), 41–47 (2011)
3. Kekre, H.B., Thepade, S.D., Sarode, T.K., Suryawanshi, V.: Image Retrieval using Texture Features extracted from GLCM, LBG and KPE. International Journal of Computer Theory and Engineering 2(5), 1793–8201 (2010)
4. Pornpanomchai, C., Supapattranon, P., Siriwisesokul, N.: Leaf and Flower Recognition System (e-Botanist). IACSIT International Journal of Engineering and Technology 3(4), 347–351 (2011)

5. Wu, S.G., Bao, F.S., Xu, E.Y., Wang, Y.-X., Chang, Y.-F., Xiang, Q.-L.: A Leaf Recognition Algorithm for Plant Classification Using Probabilistic Neural Network. In: IEEE International Symposium on Signal Processing and Information Technology. IEE explore library, pp. 11–16 (2007)
6. Tang, L., Tian, L., Steward, B.L.: Classification of Broadleaf and Grass Weeds Using Gabor Wavelets and An Artificial Neural Network. Transactions of the ASAE 46(4), 1247–1254 (2003)
7. Benčo, M., Hudec, R.: Novel Method for Color Textures Features Extraction Based on GLCM. Radio Engineering 16(4), 64–67 (2007)
8. Muralidharan, R., Chandrasekar, C.: Object Recognition using SVM-KNN based on Geometric Moment Invariant. International Journal of Computer Trends and Technology (July-August 2011)
9. Kadir, A., Nugroho, L.E., Susanto, A., Santosa, P.I.: Leaf Classification Using Shape, Color, and Texture Features. International Journal of Computer Trends and Technology, 225–230 (July- August 2011)
10. Patil, J.K., Kumar, R.: Color Feature Extraction of Tomato Leaf Diseases. International Journal of Engineering Trends and Technology 2(2-201), 72–74 (2011)
11. Kadir, A., Nugroho, L.E., Susanto, A., Santosa, P.I.: A Comparative Experiment of Several Shape Methods in Recognizing Plants. International Journal of Computer Science & Information Technology (IJCSIT) 3(3), 256–263 (2011)
12. Fiel, S., Sablatnig, R.: Automated identification of tree species from images of the bark, leaves and needles. In: 16th Computer Vision Winter Workshop Austria (2011)
13. Porebski, A., Vandenbroucke, N., Macaire, L.: Selection of Color Texture Features from Reduced Size Chromatic Co-occurrence Matrices. In: IEEE International Conference on Signal and Image Processing Applications (2009)
14. Kebapci, H., Yanikoglu, B., Unal, G.: Plant Image Retrieval Using Color, Shape and Texture Features. The Computer Journal Advance Access Published, 1–16 (2010)
15. Yang, M., Kidiyo, K., Joseph, R.: A survey of shape feature extraction techniques. In: Yin, P.-Y. (ed.) Pattern Recognition (2008)

Applying Adaptive Strategies for Website Design Improvement

Vinodani Katiyar, Kamal Kumar Srivastava, and Atul Kumar

Department of Computer Science & Engineering
SRMCEM, Lucknow (U.P). India
drvinodini@gmail.com, {2007.srivastava,atulverma16}@gmail.com

Abstract. The use of web data mining to maintain websites and improve their functionalities is an important field of study. Web site data may be unstructured to semi structured whose purpose to show the relevant data to user. This is possible only when we understand the specifics preferences that define the visitor behavior in a web site. The two predominant paradigms for finding information on the Web are navigation and search subsequently we can design a adaptive website. With the growth of World Wide Web, development of web-based technologies and the growth in web content, the structure of a website becomes more complex and web navigation becomes a critical issue to both web designers and users. In this paper we propose the method to know the significance of website by applying adaptive strategy that is dynamic map and highlights and buffering .The effect of dynamic map is apparent it can improve adaptive level of a website. Highlighting can shorten the time to find user's objective pages, and buffering pages can reduce page's response time.

Keywords: Web usage mining, Web Topology, User Navigation Pattern, adaptive strategy.

1 Introduction

As the recognition and density of word wide web (www) increases. Most Web users typically use a Web browser to navigate a Web site. They start with the home page or a Web page found through a search engine or linked from another Web site, and then follows the hyperlinks they think relevant in the starting page and the subsequent pages, until they have found the desired information in one or more pages. They may also use search facilities provided on the Web site to speed up information searching. For a Web site consisting of a very large number of Web pages and hyperlinks between them, these methods are not sufficient for users to find the desired information effectively and efficiently.

Adaptive websites are"websites that semi-automatically improve their organization and presentation by learning from user access patterns"[1].

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 857–867.
springerlink.com

The research of adaptive web site is becoming increasingly important. The adaptive technology will bring the following advantages to web users and site operators-

(i) To provide users with personalized service. Adaptive technology is able to recommend certain content to different users, according to their different hobbies and interests.

(ii) To increase system efficiency. Through log mining, we can discover the need and interest of users, optimize the pages which are in strong demand, and forecast the next page which most likely would be visited. These pages can be load into local cache, which therefore helps to balance the server load, optimize the transmission, reduce congestion, shorten the waiting time, and improve the system efficiency and service quality.

(iii) To optimize the designing of site topology according to users' previous access information, mining their access pattern can help to optimize the structure of web site and greatly enhance users' satisfaction [2].

2 Website Design Improvement

2.1 The Process of Log Mining

The adaptive website uses the web log data to to gain the user preferences. For example how many visitor visited this website, from where they are accessing, which pages are more popular etc.

Before data mining techniques are applied to web log file data, several preprocessing steps should be done in order to make web log file data ready to be mined. The process of web log mining may be divided in two parts-

1. Data Pre-treatment-It means transform the server log as the suitable form, including the data cleaning up, the user recognition and transaction recognition.
2. Use algorithms to obtain the users' visit pattern and their preference from the existing transaction set.

The following steps are used for web log mining [3]-

- Remove all the data tracked in Web logs that are useless for mining purposes e.g.: requests for graphical page content; or even requests performed by robots and Web spiders.
- According to user's IP address, divide the whole log In to independent access record sets.
- Sort the requests in every access record set according to the time submitted and then set the threshold t_w for time windows to separate the access record set. If the time between a user's adjacent page requests is shorter than t_w, the two requests are defined to be in a same user session. In the end, every session of a user makes up an access transaction.

2.2 Topology Analysis

In the design of a website, the factors like depth, connectivity, number of hyperlinks contained in the page, the frequency of being visited by users and other information plays an important role.

Initially most of the websites are designed as a tree form and then links to related nodes would be added for the convenience to users, consequently forming mesh topology. These links, though, are usually one-way, the number is relatively small and they usually have little effect on the whole topology of the site. So, the topology of site is viewed as tree structure in this paper [4]

Web log mining takes user's preference into account, combines the database of user's access interaction and the current node being visited to determine the most likely future access path. Though, only highlighting the node which would be probably visited is far from enough [5]. There are two reasons:

1. Customers visiting the site usually also want to get his relative position in the entire site. Hence it needs to extract the relative important nodes and to display to users in an appropriate manner, which would help users to aware of the environment of their position. These nodes are referred to as "landmark node". Landmark node can be manually extracted by site manager according to their experience. But this way would cost large workload, so this paper proposes an automatic algorithm to achieve this adaptive strategy.

2. Website pages are usually divided into content pages and navigation page. Obviously, some pages are not necessarily user's ultimate goal node, which means these nodes only play a navigation part in user's access affair. That is why we have tried to show only those pages to users which are useful to them [5]

2.3 Correlative Indicators

Landmark node: The nodes that are having higher connectivity to other pages, more close to home page and higher degree of users' preference. Landmark nodes are the geographical and indication nodes of the entire web pages.

Connectivity: The connectivity of a node refers to sums of the amount of other nodes which can directly visit this node and the amount of nodes which users are able to directly access to through this node[6]. The connectivity of a node may be calculated using the following equation-

$$\textit{Connectivity } C = \textit{In degree } I + \textit{Out degree } O \tag{1}$$

We have assumed website is designed using tree topology

So the connectivity of a node can be obtained by traveling the whole tree and then calculating the number of each hyperlinks of each node. In degree of a node is equivalent to the node's hyperlinks it contained, while a out degree out degree equal to the number of other nodes' hyperlinks which are directly connecting to this node. Relative connectivity of a node is calculated using following formula:

$$\textit{Relative connectivity } R_C = C / T_C \tag{2}$$

T_C is the sums of connectivity of all nodes in a web site.

Depth: Website is usually designed as hierarchical structure, which means that the page at higher level for example home page represents a higher conceptual level, while leaf nodes contain more specific content and service information. So a relative depth of a node is also an important pointer to measure its importance [6]. In this paper, the depth of a node means level of a page in the site server file system. Obviously, the node with higher depth is less important. The following formula shows

$$\text{Relative depth } R_D = 1 \text{ /Node's Depth } D \tag{3}$$

Preference: Except the structure of nodes, preference of users to nodes is one more key factor for analyzing importance of nodes. Obviously a node which is visited frequently or where users stay for a long while indicates that it has a higher degree of preference. In this paper we considered the time spent in each page visited. the time spent in each page visited is calculated by the formula $t= t_j-t_{j-1}$, where t is the time spent on the page by the user and preference is calculated by the following equation

$$\text{Node preference } R_T = T / T_A \tag{4}$$

T represents the sum of visited duration of this node and T_A is the sum of all nodes' visited duration.

Landmark coefficient: Landmark coefficient I is to measure the importance of a node:

$$\begin{aligned} \textit{Landmark coefficient } I = \textit{Relative connectivity} \times W_1 + \textit{Relative depth} \times W_2 + \\ \textit{Node preference } R_T \times W_3 \end{aligned} \tag{5}$$

Where, W_1, W_2 and W_3 stand for connectivity weight, depth weight and preferences weight respectively, and also $W_1+ W_2+W_3=1$.

The nodes with more importance are supposed to locate at a near to root, which can be used as reference coordinates by browsers. Landmark node can be screened out of nodes by dynamically setting importance threshold according to users' familiarity with the site.

3 The Strategies of Adaptive Web Site

3.1 Intelligent Recommendation

Highlight: The pages with higher probability of being visited are displayed with highlight. Site administrator can set a threshold to control the numbers of links to highlight according to real situations. Also, they can vary different colors or other means to express the degree of recommendation depending on the nodes' probability of being visited. A web page often has hundreds of links, so it would make users at a loss for searching information if unreasonable organizations of links and unclear hierarchical structure are designed [7].

Dynamic map: Another way for smart recommendation is dynamic map strategy. This strategy recommends those links that current node cannot directly access to. It is those nodes meeting a certain landmark coefficient that are recommended.

Topological analysis can be used to precisely calculate nodes' landmark coefficients and those conforming to a certain range of coefficient are candidate ones [2].

3.2 Buffering for Pre-fetching

The web pages may be cached in to server for perfecting. Pre-fetching is to put those nodes that are most probably visited to cache in advance. For example, through buffering the node the users are most likely to visit, the page can be directly fetched out from cache to apparently reduce response time [2].

3.3 Examples

Topology diagram: The following Figure1 is the topology of a small site taken as an example, whose operation system is Windows 2003 server, script language is ASP, and web server is IIS. For simplicity of understanding, names of web pages are replaced by alphabets and the extension names (.asp) are omitted as well.

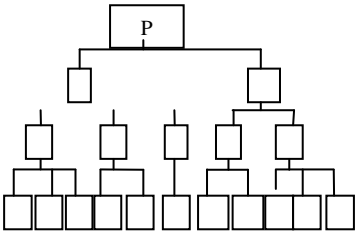


Fig. 1. An Example site topology of Azad Degree College (www.adclko.org)

Process log data in web server: This site is deployed in Windows 2003 Server environment. The log data must be cleansed including removing the track records whose http status code are 404(Not found), whose request target are *. gif (picture) and *. css (cascading style sheets), and other trashy entries. Then identify users according to clients' changing IP address. In this way, by data cleansing and users recognizing, log data within a certain period can be trimmed.

Recognize associations and predict path: Using timeout method and setting, the users' access associations are recognized as follows:

- P-R-W-R-V-V1
- P-Q-P-R-V-V2-V-V1
- P-R-W-W1-W-R-V
- P-R-V-V1-V-R-W-R-P-Q
- R-P-Q-P-R-V-V2
- P-R-V

- **Example entry of weblog of ADC**
- #Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs-version cs(User-Agent) cs(Cookie)

cs(Referer) cs-host sc-status sc-substatus sc-win32-status sc-bytes cs-bytes time-taken

- 2010-05-31 04:28:22 W3SVC4507 H-NAMADA-EWEBG 72.52.252.82 GET /Index.asp - 80 - 123.125.68.82 HTTP/1.1 Baiduspider+ (+http://www.baidu.com/search/spider.htm) - - www.adclucknow.org 200 0 0 24197 196 1640

- **Weblog** **Visiting Time Δt**
- 05:16:44 72.52.252.82 GET P.asp 200 00:01:11
- 05:16:47 72.52.252.82 GET /R.asp 200 00:00:15
- 05:22:33 72.52.252.82 GET /W.asp 200 00:02:50
- 05:16:44 72.52.252.82 GET R.asp 200 00:00:07
- 05:16:47 72.52.252.82 GET /V.asp 200 00:01:05
- 05:22:33 72.52.252.82 GET /V1.asp 200 Avg_T(V1)
- ----------
- 06:02:20 17.201.54.97 GET /P.asp 200 00:01:01
- 06:02:21 17.201.54.97 GET /Q.asp 200 00:01:01
- 06:02:26 17.201.54.97 GET /P.asp 200 00:00:04
- 06:02:30 17.201.54.97 GET /R.asp 200 00:00:03
- 06:02:30 17.201.54.97 GET /V.asp 200 00:02:06
- 06:02:31 17.201.54.97 GET /V2.asp 200 00:05:06
- 06:02:31 17.201.54.97 GET /V.asp 200 00:00:05
- 06:02:20 17.201.54.97 GET /V1.asp 200 Avg_T(V1)
- ----------
- 09:22:51 59.94.129.8 /GET /P.asp 200 00:00:15
- 09:23:06 59.94.129.8 /GET /R.asp 200 00:00:15
- 09:23:21 59.94.129.8 /GET /W.asp 200 00:01:40
- 09:25:01 59.94.129.8 /GET /W1.asp 200 00:03:15
- 09:28:16 59.94.129.8 /GET /W.asp 200 00:00:03
- 09:28:19 59.94.129.8 /GET /R.asp 200 00:00:03
- 09:28:22 59.94.129.8 /GET /V.asp 200 Avg_T(V)
- ----------
- 11:15:15 10.16.5.1 /GET /P.asp 200 00:02:51
- 11:18:06 10.16.5.1 /GET /R.asp 200 00:00:09
- 11:18:15 10.16.5.1 /GET /V.asp 200 00:00:11

Fig. 2. Web log records of Azad Degree College (www.adclko.org)

```
▪
▪       11:26:47                10.16.5.1   /GET   /V1.asp   200
        00:08:21
▪       11:26:53                10.16.5.1   /GET   /V.asp    200
        00:00:06
▪       11:27:04                10.16.5.1   /GET   /R.asp    200
        00:00:11
▪       11:31:15                10.16.5.1   /GET   /W.asp    200
        00:04:11
▪       11:31:16                10.16.5.1   /GET   /R.asp    200
        00:00:01
▪       11:31:19                10.16.5.1   /GET   /P.asp    200
        00:00:03
▪       11:31:19                10.16.5.1   /GET   /Q.asp    200
        Avg_T(Q)
▪   ----------------------------------------------------
▪       12:22:07                192.168.10.8  /GET  /R.asp   200
        00:00:20
▪       12:22:27                192.168.10.8  /GET  /P.asp   200
        00:04:52
▪       12:27:19                192.168.10.8  /GET  /Q.asp   200
        00:03:03
▪       12:30:22                192.168.10.8  /GET  /P.asp   200
        00:00:02
▪       12:30:24                192.168.10.8  /GET  /R.asp   200
        00:00:02
▪       12:30:26                192.168.10.8  /GET  /V.asp   200
        00:05:05
▪       12:30:26                192.168.10.8  /GET  /V2.asp  200
    Avg_T(V2)
▪   ----------------------------------------------------
▪       11:31:16                10.16.5.1   /GET   /P.asp    200
        00:02:51
▪       11:31:19                10.16.5.1   /GET   /R.asp    200
        00:00:09
▪       11:31:19                10.16.5.1   /GET   /V.asp    200
    Avg_T(V )
```

Fig. 2. (*continued*)

Taking node P for example, there are three associations containing P, and 9 times of accessing actions to node P, including 6 actions from P to R meaning the probability is 2/3. Therefore, the link from P to R should be highlighted or buffered in advance for shortening the response time. Node R hardly ever appeared as the last node, so R is more likely a navigation page. Taking visited duration into account can help to judge such possibility. The visited durations of R are: 15 s, 7s, 3s, 15s, 3s, 9s, 11s, 1s, 20s, 2s, 9s. C is never visited for more than 20s. So, if the threshold is set as

20s, R can be defined as navigation page, which mean visiting R is not to concern about its content, but to access other pages (such as V or W). At this time, we can use the method of dynamic map, which directly recommends the links pointing to V and W. Further, in total of 11 times of action visiting R, the probability of visiting V is 6/11, while visiting W is 3/11 and P is 2/11. So, in dynamic map, the link referring to V can be highlighted and V can be buffered as well so as to shorten the response time.

Topology analysis: Landmark coefficient is used to measure the node importance. Actually, the visiting time to the last page in an affair is hard to calculate. For a medium/large web site, the average visiting time of each page generally tends to be stable as the number of visiting increases. Therefore, We have considered, the visited duration of the last page is set to be the average of visited duration of this node. In the above example, there appear W1, W, Q and W2 as the last node. Their average visited duration is calculated as follows:

Table 1.

$Avg\_T(V1) = \sum t_i/n = 501/1=501$
$Avg\_T(V) = \sum t_i/n = 518/6=86.3$
$Avg\_T(Q) = \sum t_i/n = 244/2=122$
$Avg\_T(V2) = \sum t_i/n = 306/1=306$

Where, t_i is the time length staying at a node at the ith time, n is the total times of visiting to a node. Supposing that $W_1=W_2=W_3=1/3$, the landmark coefficients of nodes can be obtained as follows:

Table 2.

I(P)=(1/3)*2/16+(1/3)*1/1+(1/3)*790/4776=0.4301
I(Q)=(1/3)*4/16+(1/3)*1/2+(1/3)*366/4776=0.2755
I(R)= (1/3)*3/16+(1/3)*1/2+(1/3)*95/4776 =0.2358
I(S)= (1/3)*4/16+(1/3)*1/3+(1/3)*0=0.1944
I(T)= (1/3)*3/16+(1/3)*1/3+(1/3)*0=0.1736
I(U)= (1/3)*2/16+(1/3)*1/3+(1/3)*0=0.1528
I(V)=(1/3)*3/16+(1/3)*1/3+(1/3)*691/4776=0.2218
I(W)=(1/3)*4/16+(1/3)*1/3+(1/3)*524/4776=0.2310
I(V1)=(1/3)*1/16+(1/3)*1/4+(1/3)*1503/4776=2091
I(V2)=(1/3)*1/16+(1/3)*1/4+(1/3)*612/4776=0.146
I(W1)=(1/3)*1/16+(1/3)*1/4+(1/3)*195/1853=0.139
I(S1)=I(S2)=I(S3)=I(T1)=I(T2)=I(U0)=I(W2)=I(W3)=(1/3) *1/16+(1/3)*1/4+(1/3)*0=0.1042

Obviously, node Q has the highest landmark coefficients except homepage P. When users are visiting other nodes without direct physical links, dynamic map can

be introduced to recommend links pointing to Q in order to facilitate user's visit. In addition, because the landmark coefficients of V and W are similar to their father node R, when these nodes are visited, dynamic map can recommend those nodes whose landmark coefficients are close to these nodes. Also, the above analyses enlighten site designers on editing site topology. For example, with higher landmark coefficients, V and W, even V1 can be adjusted as P's direct sub-node for achieving the customer-centered site design philosophy.

4 Experimental Evaluation

4.1 Experiment Design

Adaptive strategies take account of smart highlight, dynamic map and page buffering. Here we take dynamic map for instance to analyze the effect of adaptive site. This paper compares the visit efficiency before and after adopting dynamic map using following evaluations-

(i) ***Length of visit path*** that is the number of visiting nodes. It is obvious that the shorter path means more adaptive and the higher efficiency.
(ii) ***Length of visited duration***. By dynamic map, users can save much time in navigation pages, so it would take less time to reach the destination.

4.2 Dynamic Map

According to the results of the topology analysis shown in **Table-**1 and 2 in section 3, this paper will use the dynamic map as following Table-3, where P is recommended at every node

Table 3.

Node	Strategies in Dynamic Map
P	The link pointing to Q with higher landmark coefficients is highlighted. Dynamic map provides links pointing to V and W. The link to V is highlighted.
Q	Dynamic map recommends the links directing to node R, V and W whose landmark coefficients are relatively higher. The link to V is highlighted.
R	The link to V is highlighted. Dynamic map provides links directing to Q whose landmark coefficients is high and to V1 whose landmark coefficients is close to R's.
V	Link to V1 is highlighted; Dynamic map recommend links to Q and W.
W	The links to Q and V are recommended in dynamic map.
V1	Dynamic map provides links to Q whose landmark Coefficients is high and to W whose landmark coefficients is close to V1's.
V2	The links to Q and V1 are recommended in dynamic map.

Here node R is regarded as navigation page, which means users visiting R does not need its own content. Also, those nodes where user stay for less than 20s is defined as middle node. Generally, users accessing middle node intend to obtain the links to other nodes rather than to get its contents. If dynamic map enable to directly give those posterior pages, there is no necessity for users to pass these middle node. From data information, node R, P, and W can be considered as middle nodes, so we can reach the results of contrast, which are showed as Table-4

Table 4.

User ID	Without dynamic map	With dynamic map	Rate of node reduction
72.52.252.82	P-R-W-R-V-V1；P-Q-P-R-V-V2-V-V1	P-W-V-V1；P-Q-V-V2-V1	5/14
17.201.54.97	P-R-W-W1-W-R-V；P-R-V-V1-V-R-W-R-P-Q	P-W-W1-W-V；P-V-V1-W-Q	7/17
59.94.129.8	R-P-Q-P-R-V-V2；P-R-V	R-P-Q-V-V2；P-V	3/10
Total	41	26	15/41

5 Conclusion and Future Directions

Making changes to the links of the website using dynamic maps can facilitate user's navigation of the website and minimize the time required to reach the target page. The results shown in the above table1 and table 2 shows that the effect of dynamic map is apparent. Considering the topology of middle/large sites is more complex and user's path is longer. It is no doubt that the effect would be more apparent. Consequently, such a dynamic map can improve adaptive level of a website. Highlighting can shorten the time to find user's objective pages, and buffering pages can reduce page's response time. In future it is possible to improve the way of calculating the landmark coefficient of a node, the weights of three indexes are set subjectively. Since the existence of the client cache and some pages are read from cache, not all user's visit requests are recorded in server-side's log. so the server's log cannot fully reflect the information of user's visit.

References

1. Perkowitz, M., Etzioni, O.: Adaptive Web sites: automatically synthesizing Web pages. In: Proc. of AAAI 1998, July 26-30, pp. 727–732 (1998)
2. Wang, H., Liu, X.: Adaptive Site Design Based on Web Mining and Topology. In: 2009 WRI World Congress on Computer Science and Information Engineering, CSIE, vol. 5, pp. 184–189 (2009)

3. Zhu, Z., Deng, G.: Mining Interest Association Rules in Website Based on Hidden Markov Model. In: 4th International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2008, October 12-14, pp. 1–4 (2008)
4. Deng, Y., Li, M.: Research on Web Mining and Tools. Computer Engineering and Applications (20), 92–94 (2002) (in Chinese)
5. Lu, L.: Sequential Patterns Recognition in Web Log Mining. Mini-Micro Systems 21(5), 481–483 (2000) (in Chinese)
6. Xing, D., Shen, J., Song, Q.: Discovering Preferred Browsing Paths from Web Logs. Chinese Journal of Computers 26(11), 1518–1523 (2003)
7. Valdez, F., Chignell, M.: Browsing Models for Hypermedia Databases. In: Proc. of the Human Factors Society (32nd Annual Meeting), Santa Monica, vol. 196 (1988)

WebTrovert: An AutoSuggest Search and Suggestions Implementing Recommendation System Algorithms

Akrita Agarwal[1], L. Annapoorani[2], Riya Tayal[3], and Minakshi Gujral[4]

[1] Department of Computer Science , Jaypee Institute Of Information and Technology
akrita.ag@gmail.com
[2] Department of Computer Science , Jaypee Institute Of Information and Technology
annapoorani.ln@gmail.com
[3] Department of Computer Science, Jaypee Institute Of Information and Technology
tayal.riya@gmail.com
[4] Department of Computer Science, Jaypee Institute Of Information and Technology
minakshi.gujral@jiit.ac.in

Abstract. There are hundreds of websites and apps that are struggling to find the algorithms for the perfect search to optimize the website's resources, however we have very few success stories.

We aim to build in this paper, a recommendation system, WebTrovert, which is based on practically designed algorithms. It comprises of a social networking platform holding user information and their data in the form of documents and videos.

It incorporates autosuggest search and suggestions to enhance the productivity and user friendliness of the website.

1 Introduction

In their simplest form recommender systems [1] provide a personalized and ranked lists of items by predicting what the most suitable items are, based on the user's history, preferences and constraints.

Our Recommendation System proposal is based on the approach of :

Pull and Push [3]

The main distinction between the two being that one is mainly requested (PULL) and other is mainly recommended (PUSH).

Our recommendation systems uses the hybrid filtering[1] to display the PULL and the PUSH, both of which work on the type of the user. Either the user is registered(personalised) or new(non personalised) . Here we list the various algorithms used in the paper, that perform the recommendation task making use of hybrid filtering.

- In-text Impact Index[2]
- General Popularity[1]
- Rejection Theorem[1]

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 869–878.
springerlink.com

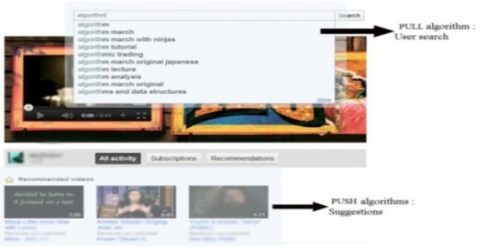


Fig. 1. A popular site screenshot illustrating the concept of PULL vs PUSH

- Item to Item Based Collaborative Filtering[4]
- Distance Similarity Matrix[1]

Table 1. The hybrid recommendation system divisional work-1

NEW USER	Content Based Filtering	Collaborative Filtering
Search (PULL)	In-text Impact Index	General Popularity
Suggestions (PUSH)	Distance Similarity Matrix	General Popularity

Table 2. The hybrid recommendation system divisional work-2

EXISTING USER	Content Based Filtering	Collaborative Filtering
Search (PULL)	In-text Impact Index Rejection Theorem	General Popularity
Suggestions (PUSH)	Distance Similarity Matrix Rejection Theorem	Item to Item Based Collaborative Filtering General Popularity

2 Brief Retrospection

WebTrovert consists of a technical social networking platform where users can signup , login , maintain profiles , upload projects, videos and documents.Videos and documents can be liked or commented upon , by users. All the content is stored in a database.

Due to the enormity of the system and database , it is impossible to be able to locate the content according to the requirement and interests .

In this paper we illustrate the various algorithms that attempt to help find other users, their content , the ratings of the uploaded content . We provide the suggestions at the bottom of the page that refresh regularly.

3 Designing the Search System

Websites must implement the search system efficiently for a user friendly browsing. Here we implement a search system for an existing user , which implements the recommender systems using an autocomplete that adds real time computing to the search.

Hence forth we call the system "WebTrovert". Basic system is illustrated in fig 2, pseudocodes shown below.

Pseudocode to design the search box :

search: A search textbox which calls lookup() onkeyup(id=inputstring)

The function lookup() is a function which can use jquery to call the php file getdatabase.php which fetches the database :

function lookup(inputString)
if (length of 'inputstring' equal to 0) Hide the suggestion box.
else
call getdatabase.php with 'inputString' and 'queryString'
(string entered in the search textbox)concatenated to the url
function(data)
if lengthOfData is greater than 0 show suggestions

Fig. 2. Basic webTrovert search system

Now in the next section we describe the algorithms that work behind the search box to aid the autosuggest .

3.1 In-text Impact Index Factor (ItIF)

The ItIF is the basic algorithm used by all search systems that displays the information as per the user query. It may display infinite data, most of which may not be useful or relevant to the user context.

The ItIF result set mainly acts as the input set for other algorithms, thus refining the otherwise infinite search result set. Fig 4 illustrates the WebTrovert search using ItIF , coded below using php .

In getdatabase.php(pseudocode given here), we capture the value of querystring sent from the function lookup , and use it to fetch data from database using mysql query:

If 'queryString' is set
If length of 'queryString' greater than 0
*query =SELECT * FROM iteminfo WHERE itemname LIKE '%$queryString%' OR docdesc LIKE '%$queryString%') LIMIT 30"*
if ('query' is valid) fetch query results one by one
and display this data in a drop down list //these are the suggestions

Fig. 3. Php query to implement ItIF

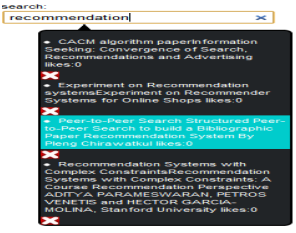


Fig. 4. WebTrovert Autosuggest depicting ItIf

Note: For the following algorithms , we only update the mysql query .

3.2 General Popularity Algorithm

General popularity algorithm displays the result set of ItIf in decreasing order of their popularity. Here we measure the popularity factor by comparing the "no of views" of each result of ItIF. Fig 5 Illustrates the working of ItIf and general popularity Fig 5 illustrates the WebTrovert search using ItIF and General Popularity . The corresponding mysql query is given below below .

$query = $db->query("SELECT * FROM itemdatabase WHERE (itemname LIKE '%$valueintextbox%' OR itemdescription LIKE '%$valueintextbox%') ORDER BY views DESC ;LIMIT 30");

Fig. 5. WebTrovert Autosuggest search depicting General Popularity and ItIF

3.3 Rejection Theorem

The rejection theorem works on the principle that user may not like some results appearing in his "Search result" and thus may wish to remove the results.

Thus , We remove the search result from the user suggestions once it is rejected, but instead of completely rejecting it, we append it to the end of our suggestion list, thus making it possible to be displayed in case of specific result based search or requirement. Fig 6 illustrates the working on rejection theorem , when the user rejects a result,

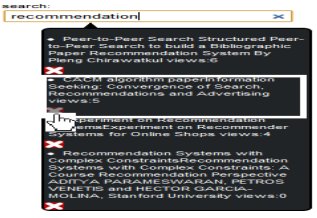


Fig. 6. WebTrovert Autosuggest search depicting General Popularity and ItIF; user rejecting a result by clicking on the "cross"

The following code is executed in the while loop of the code given in Fig 3:

Fetch query result one by one

On selecting a result fill the textbox with the selected value

On clicking cross function removethisdata() called with username and itemurl as arguments

Once a "cross" is clicked , the function removedata() makes an entry into the removeitem table , thus blacklisting the current item , for the current user . the query for entering into removedoc :

sql="INSERT INTO removeitem(username,itemurl) VALUES('$user','$itemurl')";

Thus ,query for fetching the data in getitem.php is modified to be :

*query =SELECT * FROM iteminfo WHERE (itemname LIKE '%$queryString%' OR itemdesc LIKE '%$queryString%') AND itemurl NOT IN (SELECT itemurl FROM removeitem WHERE username = 'current username') ORDER BY views DESC LIMIT 30");*

Thus finally the suggestions appear as given in Fig 7.

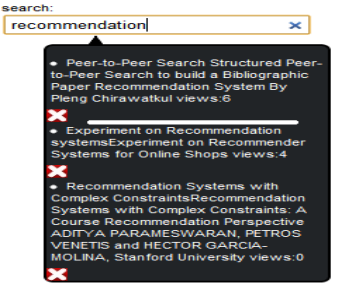


Fig. 7. WebTrovert Autosuggest search depicting General Popularity, ItIF and Rejection Theorem

4 Designing the Suggestion System

Websites implement the search system efficiently for a user friendly browsing. however analyzing the user interests in not just limited to autosuggest search results, it

gives the system an ability to intelligently suggest to users , content they might like, irrespective of what they might actually be looking for at that moment. This is called the suggestion system which is implemented via the following 2 algorithms.

4.1 Item to Item Based Collaborative Filtering

For each of the user's liked items, the algorithm attempts to find similar items. It then aggregates the similar items and recommends them.

Everytime user likes an item, given in the following code, the correlation matrix and cos similarity matrix are updated. The matrices are elaborated in the following sub section.

mapping the « like » option :

a like button which on clicking triggers viditemcorr(videoid)

4.1.1 Item to Item Based Correlation Matrix

Suppose we have the following details

Table 3. User to item relation example

Documents liked →

user	Piano (P)	Guitar (G)	Balloon (B)	Table (T)
Harry	Yes	Yes	No	No
Ron	Yes	No	No	Yes
Herm	No	Yes	Yes	No

Here we have 4 items liked by 3 users . Now we assume :
1 -> yes ; 0 -> no
And each item corresponds to a vector . Thus
Piano -> vector P(1,1,0)
Guitar - > vector G(1,0,1)
Balloon -> vector B(0,0,1)
Table -> vector T(0,1,0)
MySql query (To generate the item to item correlation matrix):
UPDATE vidlikes SET likes = likes + 1 WHERE (username = 'current user' AND vidid='$_GET[vid]') ;

It can measure the similarity of twoitems, A and B, in various ways; a common method is to measure the cosine of the angle between the two vectors:

$$similarity(\vec{A},\vec{B}) = \cos(\vec{A},\vec{B}) = \frac{\vec{A} \bullet \vec{B}}{\|\vec{A}\| * \|\vec{B}\|}$$

$$[P\ G] = \frac{P \cdot G}{\|P\|\ \|G\|}$$

$$= \frac{(1\ 1\ 0).(1\ 0\ 1)}{\|1\ 1\ 0\|\ \|1\ 0\ 1\|}$$

$$= ½$$

$$[P\ B] = \frac{P \cdot B}{\|P\|\ \|B\|}$$

$$= \frac{(1\ 1\ 0).(0\ 0\ 1)}{\|1\ 1\ 0\|\ \|0\ 1\ 0\|}$$

$$= 0$$

$[P\ T] = 1/\sqrt{2}$

Pseudocode code (to calculate the cos similarity values) :

```
Set i=1
Loop till  i<=no ofusers
        sum = sum +svector[i]*dvector[i]
            if (square root(scount)*square root(dcount))
                  sum = sum/((scount)*square root(dcount));
           sendme = (float)sum;
                              Increment i by 1
```

Where $svector → source vector e.g. as we used Piano above
$dvector → destination vector e.g. as we used Guitar , Balloon above

Here [P G] > [P T] > [P B]

Therefore if a user views Piano(P) then he/she will be recommended first Guitar(G) then the Table(T).

But since [P B] = 0, thus Balloon(B) will not be recommended .

From this we can make the Cos Similarity Matrix for the liked items.

4.1.2 Cos Similarity Matrix

Table 4. Item to item table

Cos of liked items →

Items	P	G	B	T
P	-	½	0	$1/\sqrt{2}$
G		-	$1/\sqrt{2}$	0
B			-	0
T				-

Here we fill only the upper triangle of the matrix .

Since similarity(A,B) = Similarity(B,A)

Here on Y axis we have the items and corresponding to them the recommendation preference of the other videos .

MySql query (to generate the cos similarity matrix):
INSERT INTO vidcos VALUES('$sourceid','$destid','$sendme') ;

So if a user views Piano(P), then his Recommendations are in order :
Guitar(G) > Table(T)

4.2 Distance Similarity Matrix

This algorithm works on the basis that if two documents are very similar to each other , then if one is recommended , then the other can be recommended with a conviction that it is relevant as well .To calculate the distance between two items, distance similarity uses the Levenshtein distance[2] to compute the distance ratio(DR) between items , taken 2 at a time . A levenshtein function works as follows :

Let string 1 → helloworld , string 2 → hell
Length(String 1)> length(string 2)
Thus len = length(string 1) i.e. 10.

Now 6 changes are required in string 2 to make it equivalent to string 1.
Thus levenshtein distance : 6
And distance ratio : 6/10 = 0.6
Similarly , Let String 3 → hellwor
Now levenshtein distance : 2 and distance ratio : 2/10 = 0.2
Since DR(string(1,2)) > DR(string(1,3))
2 is at a larger distance to 1 than 3. Thus 3 is distance similar to 1 than 2.
Thus the recommendation outcome of String 1 will be in the order :
String 3 > String 2

MySql function call :
SELECT distance('$string1','$string2') AS lev" ;

MySql function definition :

```
Function distance(s1,s2)
  DECLARE s1_len, s2_len, max_len INT;
  SET s1_len = LENGTH(s1), s2_len = LENGTH(s2);
  IF s1_len > s2_len THEN
    SET max_len = s1_len;
  ELSE
  SET max_len = s2_len;
  END IF;
  RETURN ROUND((1 - LEVENSHTEIN(s1, s2) / max_len) * 100);
```

This function returns the distance ratio for the 2 corresponding strings , we can insert them into a table.

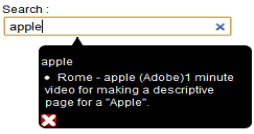


Fig. 8. Autosuggest search depicting the cumulative results of the PULL

Suggestion result of apple is an item called : rome apple

Fig. 9. WebTrovert suggestion system depicting Item to item Collaborative Filtering and Distance Similarity Matrix

We see that our initial search was « apple » however we are suggested items that do not contain the word « apple » but may be equally relevant . Here the distance ratio helps compute the relation between the items on basis of their similarity.Thus we have our suggestions.

5 Challenges Faced

Recommendation system is a major success on any platform where there are ample datasets available . However , a limited database generally leads to an inefficient performance . In such scenarios , basic analytic algorithms like ItIF are very useful.

Other problems include changing user preferences, changing data, which have been handled well in item to item based collaborative filtering technique.

However item to item filtering may face overspecialization drawbacks, in case of application on certain unpredictable items.

6 Conclusion

Search system in a website can be improved upon , using a lot of other algorithms as well . however , our WebTrovert system uses the above mentioned 5 algorithms efficiently. practical results are generated with minimization of coldstart, overspecialization and infinite dataset issues and a lot of emphasis on user interests and general trends of query.

Despite efforts by many websites , So far only a handful of companies have really gotten recommendations to a high level of user satisfaction - Amazon, Netflix (although of course they are looking for a 10% improvement on their algorithm), Google are some names that spring to mind.

But for those select few success stories, there are hundreds of other websites and apps that are still struggling to find the magic formula for recommending new products or content to their users .

Thus in the current scenario , WebTrovert aims at providing a solution to the effort of developing an efficient recommendation system.

References

[1] Agarwal, A., Annapoorani, L., Tayal, R., Gujral, M.: Recommendation systems : A practical approach (Yet To Be Published)

[2] Gipp, B., Beel, J., Hentschel, C.: Scienstein: A Research Paper Recommender System. In: Proceedings of the International Conference on Emerging Trends in Computing (ICETiC 2009), pp. 309–315 (2009)

[3] Garcia-Molina, H., Koutrika, G., Parameswaran, A.: Information Seeking: Convergence of Search, Recommendations and Advertising. ACM Transactions on Information Systems, Stanford University

[4] Linden, G., Smith, B., York, J.: Amazon.com recommendations: Item-to-item collaborative filtering. IEEE Internet Computing (January 2003)

[5] Rashotte, L.: Social influence. In: Ritzer, G. (ed.) Blackwell Encyclopedia of Sociology, pp. 4426–4429 (2007)

[6] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.: GroupLens: An open architecture for collaborative filtering of Netnews. In: Proc. ACM Conference on Computer Supported Cooperative Work. ACM Press, Chapel Hill, North Carolina, United States(1994)

A Study of the Interval Availability and Its Impact on SLAs Risk

Andres J. Gonzalez and Bjarne E. Helvik

Centre for Quantifiable Quality of Service in Communication Systems*,
Norwegian University of Science and Technology, Trondheim, Norway
{andresgm,bjarne}@q2s.ntnu.no

Abstract. The obligations that telecommunications providers have with their customers are nowadays clearly specified in SLA contracts. The offered availability during the contract period is one of the most relevant variables in SLAs. Modeling accurately the transient solution posed by the need of considering the interval availability is still an open challenge. A common policy taken to make simpler models is the use of steady state assumptions. Nevertheless, this simplification may put on risk the contract fulfillment as stochastic variations of the measured availability are significant over a typical contract period. This paper makes a theoretical study of the interval availability and propose an approximation to evaluate the cumulative downtime distribution of a system component using renewal theory. We study the evolution of the distribution of the interval availability with the increase of the observation period i.e., duration of the contract, and show its respective impact in the SLA success probability. In addition, using the approximation proposed, we analyze numerically the behavior of the cumulative downtime distribution and the SLA risk under processes that do not follow Markovian assumptions.

1 Introduction

Real world networks are not fault free. A single failure has economic and reputation impact to the operator and incalculable consequences to the customers through the affected services. A common policy to handle this issue is the stipulation of the availability to be guaranteed in a business contract known as Service Level Agreement SLA. The promised availability must be commercially competitive and it must fit the customer needs. In addition, it may imply huge costs in terms of the price of high reliable equipment and the penalties associated with the violation of the agreement. Therefore the selection of the availability to be promised requires an accurate analysis. However, it is difficult to define due to the computational challenge posed by the transient solution and the stochastic variations of the interval availability. The use of steady state assumptions is a common policy taken to simplify the analysis. Nevertheless, this simplification

* "Centre for Quantifiable Quality of Service in Communication Systems, Centre of Excellence" appointed by The Research Council of Norway, funded by the Research Council, NTNU, UNINETT and Telenor. http://www.q2s.ntnu.no

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 879–890.
springerlink.com

may put on risk the contract fulfillment as stochastic variations of the measured availability are significant over a typical contract period.

Our study is focused on network components that can be operational (up) or not operational (down) that are modeled as a two-state system. We start by proposing a numerical method to estimate the cumulative downtime distribution of individual network comopents. The definition of the cumulative downtime probability density function PDF in a component with general distributed failure and repair processes during a finite interval has been an open challenge for a long time. The formulation of this problem was first addressed by Takács [11] and confirmed by Muth [9] using different methods. They obtained a solution in terms of the convolution of the cumulative distribution function CDF of the failure/repair processes. However, an explicit solution is only given for exponentially distributed up/down times, due to the complexity posed by the n-fold convolution of CDFs with other kinds of distributions. Takács also proposed a tractable solution for general distributions, assuming long intervals ($t \to \infty$). On the other hand, Funaky et al. [1] proposed a good approximation assuming short intervals. Given that a typical SLA duration does not fit into the two mentioned solutions, we propose an approximation to calculate the distribution of the cumulated downtime in a SLA context, i.e., a finite contract-interval that last for several months.

The probability that a network operator *meets / do-not-meet* the contracted interval availability α will be referred as *SLA success probability / risk*. This concept was first raised for general systems in [5] by Goyal and Tantawi. They observed that if the promised availability is larger than the *steady state availability* (A), the success probability decreases continuously. However, they also showed that there is a considerable risk even when $\alpha < A$. They provide a numerical method to compute risk assuming Markovian failure and repair distributions. For the case of non-markovian and complex composed systems previous works such as [3],[10] and [4] use simulation tools for the assessment.

This paper proposes a numerical method to characterize the entire distribution of the interval availability in network components that have failure/repair processes Weibull, gamma and negatively exponentially distributed. In addition the evolution of the interval availability with the duration of the contract is studied and the effects of the shape parameter on the SLA success probability are shown.

This paper is organized as follows. In Section II, we present our approach to calculate the two-state component cumulative down time during a finite interval. Section III describes the evolution of the distribution of the interval availability and studies the effect of the shape parameter in the SLA risk. Finally, Section IV concludes the paper.

2 Distribution of the Cumulative Downtime during a Finite Interval

The operational status (up/down) of a network component can be modeled as a two-state system. The right modeling of such system is crucial in order to

evaluate the unavailability and its associated penalty. In this section, we present a method to evaluate numerically the distribution of the accumulated downtime of a two-state system with general distributed ud/down times.

The network component state as a function of time can be modeled by a random process $O(T)$ defined as follows:

$$O(T) = \begin{cases} 1 & \text{If the component is working at time } T. \\ 0 & \text{Otherwise.} \end{cases} \tag{1}$$

The interval availability $\hat{A}(\tau)$ is a stochastic variable that measures the time that a network component has been working during τ.

$$\hat{A}(\tau) = \frac{1}{\tau} \int_0^\tau O(T)\, dT. \tag{2}$$

The network component behavior can be described as a sequence of failure (g) and repair (h) processes, i.e., $O(T)$ may be modeled as an alternating renewal process.

The accumulated down time over τ $t(\tau)$ is associated with $\hat{A}(\tau)$ as: $t(\tau) = \tau[1-\hat{A}(\tau)]$. $\Omega(\tau, t)$ and $\omega(\tau, t)$ are defined as the CDF and the PDF of $t(\tau)$ respectively. A general expression for $\Omega(\tau, t)$ was derived by Takács in [11] as follows

$$\Omega(\tau, t) = \sum_{n=0}^{\infty} H_n(t)[G_n(\tau - t) - G_{n+1}(\tau - t)] \tag{3}$$

where the failure and repair processes are described by i.i.d. up and down times with CDF $G(t)$ and $H(t)$ respectively, and the subindex n represents the $n-$fold Stieltjes convolution of a given function.

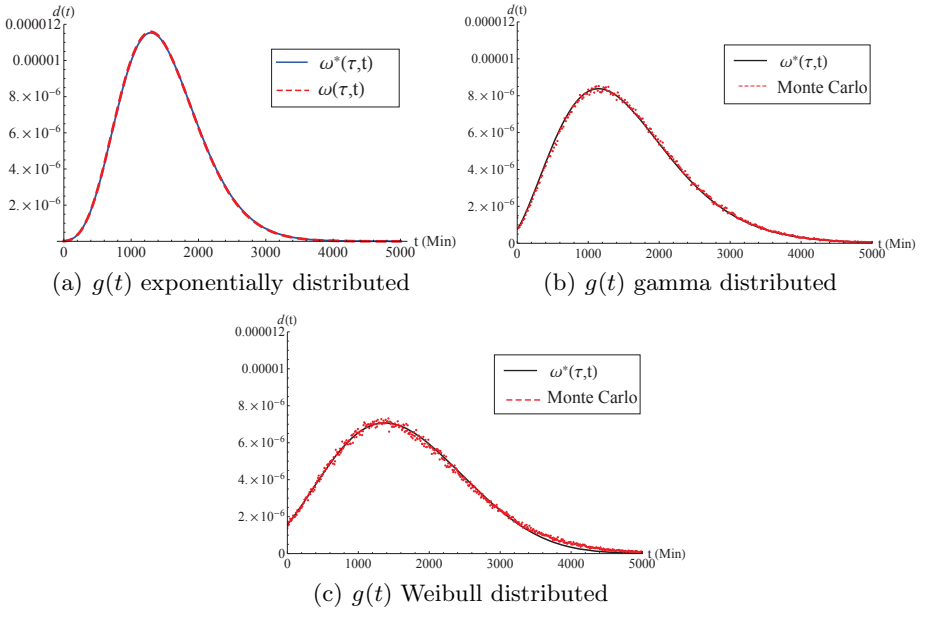
Equation (3) characterizes a problem with general distributions. However, it is difficult to compute for specific failure and repair processes due to the complexity posed by the $n-$fold convolution of general distributed CDFs. In [1], $\Omega(\tau, t)$ is approximated for general distributions assuming short intervals. Nevertheless, this result cannot be applied in our problem given that the duration of a SLA is typically of several months.

For the case of failure and repair processes exponentially distributed, a complete result was obtained by Takács as

$$\Omega(\tau, t) = e^{-\lambda(\tau - t)} \Bigg[1 + (\lambda\mu(\tau - t))^{0.5} \int_0^t e^{-\mu y} y^{0.5} I_1(2(\lambda\mu(\tau - t)y)^{0.5}) dy \Bigg] \tag{4}$$

where λ and μ are the respective failure and repair rates and I_1 is the Bessel function of order 1.

Some studies (e.g., [8], [2]) have shown that the Weibull and gamma distributions are representative to model real failure and repair processes. In the



(a) $g(t)$ exponentially distributed

(b) $g(t)$ gamma distributed

(c) $g(t)$ Weibull distributed

Fig. 1. Approximation of the PDF cumulative downtime

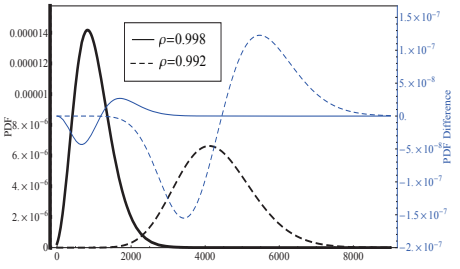
literature there is not an explicit expression that describes $\Omega(\tau, t)$ for these two distributions. In this paper, we suggest an approximation using renewal theory, making $\Omega(\tau, t)$ and $\omega(\tau, t)$ tractable and sufficiently accurate.

Assuming n down events during τ, the duration of each down period is assumed independent and identically distributed $h(t)$. The PDF of the total cumulated downtime is given by the n-fold convolution $h_n(t)$. Hence, if the probability of n down events during τ $P(N(\tau) = n)$ is known, the problem can be solved.

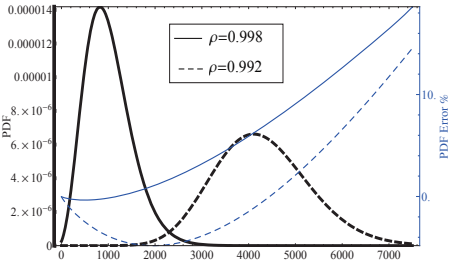
Individual network components in backbone operational networks are highly reliable with mean time to failure in the order of months and mean time to repair in the order of minutes to hours, obtaining steady state availabilities ρ usually larger than 0.999 (See for instance [2]). In this context, one can assume that the downtime duration is very small compared to the uptime. Therefore, we can approximate the number of down events during τ considering only the number of renewals of the failure process.

This approximation overcomes the complexity of (3) by dividing the problem in two. First by finding the n-fold convolution of only the PDF of the downtime, and second by obtaining the number of down events ruled only by the uptime distribution. For this, renewal theory and counting models may be used. The approximated PDF of the total cumulated downtime is given as

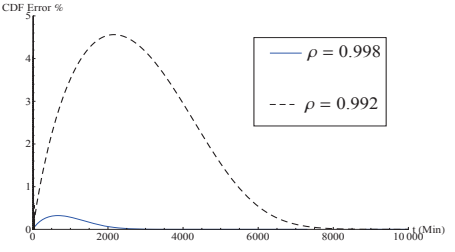
$$\omega^*(\tau, t) = \sum_{n=0}^{\infty} P(N(\tau) = n) h_n(t) \tag{5}$$



(a) **Left axis**: $\omega(\tau, t)$ density. **Right axis**: PDFs difference $(\omega^*(\tau, t) - \omega(\tau, t))$



(b) **Left axis**: $\omega(\tau, t)$ density. **Right axis**: PDF percent error $(1 - (\omega(\tau, t)/\omega^*(\tau, t)) \times 100\%)$



(c) CDF percent error $(1 - (\Omega(\tau, t)/\Omega^*(\tau, t)) \times 100\%)$

Fig. 2. Approximation error

The probability $P(N(\tau) = n)$ of n renewals during τ for exponentially distributed uptimes follows the Poisson distribution. The n-fold convolution of an exponential function can be easily obtained applying Laplace transform.

In [12], Winkelmann defines a count-data model that computes $P(N(\tau) = n)$ when the times are independent and identically gamma distributed:

$$P(N(\tau) = n) = G^i(\beta_g n, \theta\tau) - G^i(\beta_g n + \beta_g, \theta\tau) \tag{6}$$

where β_g and θ are the shape and scale parameter of the gamma distribution, respectively, and G^i is the incomplete gamma function. The n-fold convolution of gamma distributed downtimes is obtained straight forward by the Laplace transform.

Finally, the Weibull distribution poses a bigger challenge, since its Laplace transform is analytically intractable. However, for the case of Weibull distributed uptimes, $P(N(\tau) = n)$ can be obtained by expanding the Weibull function in Taylor series. In [7] for a renewal Weibull process with shape parameter β_W and scale parameter η equal to one, $P(N(\tau) = n)$ is defined as

$$P(N(\tau) = n) = \sum_{s=n}^{\infty} (-1)^{s+n} \frac{\tau^{\beta_W s}}{s!} \frac{b_k(s)}{\gamma(s)} \tag{7}$$

where $\gamma(s) = b_0(s) = \Gamma(\beta_W s + 1)/\Gamma(s+1)$ and $b_{k+1}(s) = \sum_{w=k}^{s-1} b_k(w)\gamma(s-w)$.

The convolution of Weibull distributed downtimes is approximated in [6] using the Saddle Point Approximation.

In order to see the accuracy of our approximation, Fig. 1 compares $\omega^*(\tau, t)$ with $\omega(\tau, t)$ on network components with three different uptime distributions with the same expected value. Fig. 1(a) shows a network component with exponentially distributed uptimes with $\lambda = 1/30$ days ($E(g(t)) = 30$ days) and exponentially distributed downtimes with $\mu = 1/2$ hours. $\omega(\tau, t)$ was obtained using (4).

Fig. 1(b) compares $\omega^*(\tau, t)$ with a Monte Carlo simulation when network component uptimes are gamma distributed with shape parameter $\beta_g = 0.5$, scale parameter $\theta = 60$ days ($E(g(t)) = 30$ days) and exponentially distributed downtimes with $\mu = 1/2$ hours. Finally, Fig. 1(c) compares our approximation with a Monte Carlo simulation in a network component with Weibull distributed uptimes with shape parameter $\beta_W = 0.5$, scale parameter $\eta = 15$ days ($E(g(t)) = 30$ days) and exponentially distributed downtimes with $\mu = 1/2$ hours. From Fig. 1 not only the accuracy of the approximation can be appreciated, but also the variance of $\omega(\tau, t)$, which becomes bigger when the failure processes are not exponentially distributed but Weibull or gamma distributed with shape parameters shorter than one.

In order to estimate the magnitude of the error posed by our approximation, we use three different methods. First we evaluate the PDF difference $(\omega^*(\tau, t) - \omega(\tau, t))$.

Second the percent error between PDFs $(1 - (\omega(\tau, t)/\omega^*(\tau, t)) \times 100\%)$ and finally the percent error between CDFs $(1 - (\Omega(\tau, t)/\Omega^*(\tau, t)) \times 100\%)$. We use exponentially distributed processes, given that (4) can be used as reference.

Fig. 2 presents the results of applying these three methods in two different network components with ρ =0.998 and ρ =0.992, respectively. Fig. 2(a) illustrates that $(\omega^*(\tau, t) - \omega(\tau, t))$ is approximately three and two order of magnitude smaller than $\omega(\tau, t)$ in the network component with ρ =0.998 and ρ =0.992, respectively. Fig. 2(b) shows initially a negative error that becomes zero when the cumulated downtime is equal to $E[\omega(\tau, t)]$ and it becomes positive with monotonic increase for $t > E[\omega(\tau, t)]$. However, when the PDF error becomes considerable, the remaining probability mass is small $(1 - \Omega(\tau, t) \to 0)$. In order to illustrate this better, Fig. 2(c) considers the remaining mass probability by estimating the error directly in the CDF.

Peak CDF error equal to 1% and 5% are obtained for network components with ρ equal to 0.996 and 0.9915, respectively. The presented results show that our approximation will work well in backbone networks scenarios and general systems with mean time to failure in the order of months and mean time to repair in the order of minutes to hours.

3 SLA Success Probability

The interval availability becomes fundamental in business scenarios where a clear specification of the offered availability during a contract period has to be defined.

Previous works try to assess the availability to be promised using simulation techniques. In this paper, we make use of the numerical results obtained and presented in the previous section in order to present the behavior of the interval availability in different scenarios, using numerical methods. In this section, first we will define the success and risk of an SLA. Second, the evolution of the interval availability with increasing τ will be presented. Finally the effect of the shape parameter of failure processes will be analyzed.

When a SLA is defined, the provider promises an availability guarantee α for a given period τ (the duration of the contract). Under this scenario, to know the probability that the availability after some observation period τ will be larger than or equal to the defined guarantee is crucial. The Success Probability is defined as follows:

$$S(\tau, \alpha) = Pr[\hat{A}(\tau) \geq \alpha] \tag{8}$$

Additionally, the *risk* will be defined as the probability that the specified availability α will not be met, which can be expressed as $1 - S(\tau, \alpha)$. Fig. 3 shows the general shape of the PDF of the interval availability and how the risk and the success of the SLA can be estimated from this information.

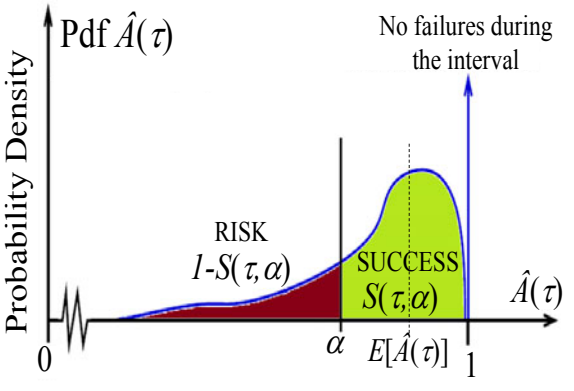


Fig. 3. Interval Availability (General Shape)

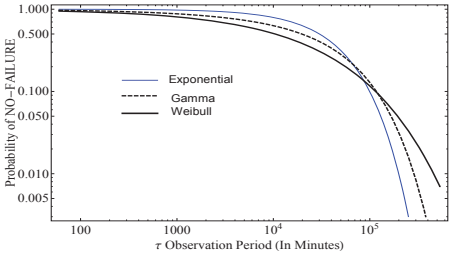
3.1 Interval Availability Evolution

The interval availability usually is modeled using simulation techniques. In this subsection, we will model accurately the shape of the interval availability, using the numerical methods described in section 2.

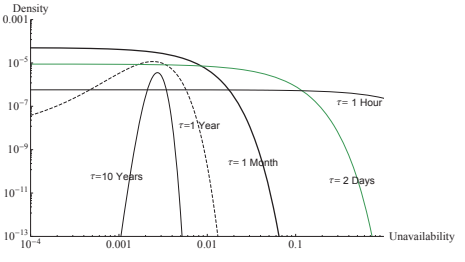
Fig. 3 shows that the density of the interval availability includes a Dirac delta function that represents the probability of no failure during the interval τ. This subsection shows explicitly the dimension of this probability i.e., the magnitude of the integral of the impulse. In section 2 was presented how to

compute the number of n down events during τ. Therefore we obtain the probability of no failures by computing $P(N(\tau) = 0)$. For the case of exponential failure processes this probability is reduced to a negative exponential function with parameter λ. For the case of gamma and Weibull distributed failure processes the probability of no failure will be computed using expressions (6) and (7) respectively.

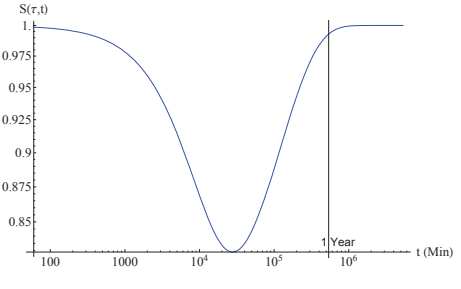
Fig. 4(a) shows the results obtained for the probability of no failure for three different failure processes. First, negatively exponentially distributed failure processes with intensity $\lambda = 1/30$ days. Second, gamma distributed failure processes with shape parameter $\beta_g = 0.5$, scale parameter $\theta = 60$ days . Finally Weibull distributed failure processes with shape parameter $\beta_W = 0.5$, scale parameter $\eta = 15$ days. The expected value for all the cases is the same ($E(g(t)) = 30$ days). We observe that for observations periods approximately shorter than 2 months ($8x10^4$ minutes) the Weibull and the gamma distribution present a higher reduction with increasing observation period τ. However, for observations periods larger than 2 months, the probability of no failure decrease faster under negatively exponentially distributed failures.



(a) Probability of NO-FAILURES during τ



(b) Interval Availability for different τ



(c) SLA success probability

Fig. 4. Evolution of the Interval Availability with τ

With the magnitude of the integral of the Dirac delta function defined, the next step is to study the rest of the distribution where interval availability values shorter than one are considered. For this, we use expression (4) in order to evaluate the interval availability in a component with negatively exponentially distributed uptimes with $\lambda = 1/30$ days and exponentially distributed downtimes with $\mu = 1/2$ hours.

We select five different values for the observation period in order to be able to describe the evolution of the interval availability with τ. Fig. 4(b) shows the obtained results using unavailability in the horizontal axes given that it offers a more illustrative presentation using a logarithmic scale. When the observation period is very short e.g. 1 hour (60 minutes), the density appears almost uniformly distributed being dominant the probability of no failure presented in Fig. 4(a) which is 0.9986. The next selected τ is two days (2880 minutes) in this case the probability of no failure is 0.9356. In addition the probability of high unavailability values i.e., bigger than 0.1 start to be strongly reduced. When the observation period is equal to 1 month (43200 minutes) the probability of no failure is reduced to 0.369 and the probability of having unavailability values higher than 0.1 and 0.01 becomes negligible and considerably reduced respectively. When τ is equal to 1 year (525000 minutes) the probability of no failure becomes very small ($5.4\text{x}10^{-6}$). In addition the probability mass start to be concentrated near to the expected interval availability value ($E[\hat{A}(\tau)]$). Finally when the observation period is very large i.e., 10 years (5250000 minutes), the interval availability present a distribution close to normal as was mentioned by Takács in [1] and the probability of no failure becomes negligible with a value of $2.4\text{x}10^{-53}$

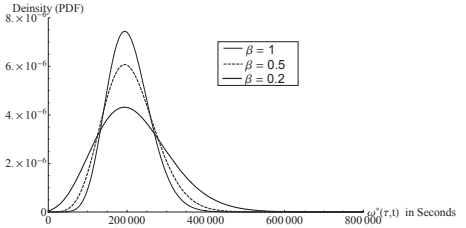
Finally for the shake of illustration, Fig. 4(c) presents the shape of the success probability. This information agrees with the results shown in [5] and [3]. The information and analysis made from Fig. 4(b) combined with Fig. 4(c) provide a better understanding of the shape and stochastic variations of $S(\tau, \alpha)$.

3.2 The Effect of the Shape Parameter

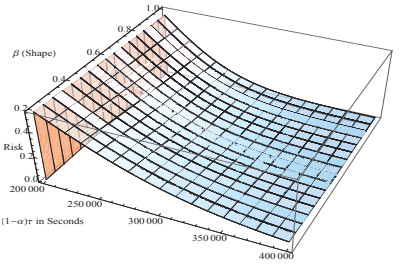
Measurements of operational systems show higher occurrence of very short and long system uptimes than what is properly described by a negative exponentially distribution, e.g. [8], [2]. They have found that Weibull and gamma with shape parameters less than one are more representative to model the behavior of uptimes real systems.

We use the results obtained in section 2 to show the effect of the uptime shape parameter in the SLA success-probability/risk.

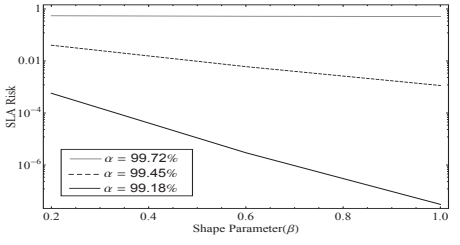
Fig. 5(a) shows the probability distribution of the cumulative downtime for an observation period of 18 months in a two-state system with negatively exponentially distributed downtimes with $\mu = 1/2$ hours and three different gamma distributed uptimes with the same average uptime duration of 1 month (the steady state availability for all these systems is 0.997230). One can observe that when β_g is equal to 1 (negatively exponentially distributed uptimes) the distribution of the cumulative downtime appears approximately symmetrically distributed



(a) $\omega^*(\tau, t)$



(b) 3D-Risk under different values of α and β



(c) Risk under different values of α and β

Fig. 5. The effect of β in $\omega(\tau, t)$ and $S(\tau, \alpha)$

around the expected value ($E[\omega(\tau, t)]$). With the reduction of the shape parameter the stochastic behavior of uptime duration becomes more bursty presenting shorter and longer times. This property produces an increase in the variance of $\omega(\tau, t)$ and a loss in the symmetry around the expected value.

When an availability promise α is stipulated in an SLA, it means that the cumulative downtime has to be shorter than $(1 - \alpha)\tau$. Therefore, an alternative notation for the risk would be:

$$\text{RISK}(\tau, \alpha) = \int_{(1-\alpha)\tau}^{\infty} \omega(\tau, t) dt. \tag{9}$$

Fig. 5(b) shows the risk values obtained after applying equation (9) in several two-state systems with the same expected values that the cases illustrated in Fig. 5(a) and a fixed observation period of 18 months. We use different values of $(1 - \alpha)\tau$ and shape parameters β_g from 0.2 to 1. Fig. 5(b) illustrates that the risk can be reduced considerably if the SLA availability promise is shorter than the expected interval availability for all the evaluated shape parameters and that the shape parameter affects in an inconvenient way the dimension of the risk for all values of α.

Finally Fig. 5(c) shows specific cuts from Fig. 5(b) using three different points (α=0.9972, α=0.9945 and α=0.9918) and studying the effects of the different

shape parameters on the risk. As observed before, the shorter the shape parameter value, the higher the magnitude of the risk. However, the farther the promise is form the steady state availability the higher the difference in orders of magnitude produced by the shape parameter.

4 Conclusion

We show that the cumulative downtime distribution of a two-state system can be accurately approximated in a backbone network context, using counting models and feasible PDF's Laplace transforms.

Pervious works have study the effects of the interval availability using simulation or via numerical methods by taking Markovian assumptions. This paper complements previous works offering a numerical method to analyze non Markovian systems.

The detailed analysis of the distribution of the interval availability helps to understand better the implications of sign availability promises in SLAs. We observe that for short observation periods the interval availability presents a dominant density concentrated in the probability of no failure and that can be modeled using Dirac delta function. For the rest of the availability values the probability density is almost uniformly distributed. When the observation period start to increase the probability of no failure and the probability of permanent failure (system always down) start to be considerably reduced and this density star to be concentrated around the expected interval availability value. Finally for very long observations periods a behavior similar to normal is obtained.

With the reduction of the shape parameter for values below one (bursty uptimes) the risk increase considerably. In addition, the smaller the promised availability is, the higher the difference produced by the shape parameters, reaching differences of several orders of magnitude.

The assumption of steady state conditions may simplify the modeling of several dependability problems. However, as this paper shows, different stochastic variations may have a huge impact in the success of an SLA.

References

1. Funaki, K., Yoshimoto, K.: Distribution of total uptime during a given time interval. IEEE Transactions on Reliability 43(3), 489–492 (1994), doi:10.1109/24.326451
2. Gonzalez, A., Helvik, B.: Analysis of failures characteristics in the UNINETT IP backbone network. In: IEEE 7th International Symposium on Frontiers in Networking with Applications, FINA 2011 (March 2011)
3. Gonzalez, A., Helvik, B.: Guaranteeing service availability in SLAs; a study of the risk associated with contract period and failure process. IEEE Latin America Transactions (October 2010), doi:10.1109/LATINCOM.2009.5304893
4. Gonzalez, A., Helvik, B.: Guaranteeing service availability in slas on networks with non independent failures. In: IEEE-IFIP International Workshop on Design of Reliable Communication Networks, DRCN (October 2011)

5. Goyal, A., Tantawi, A.: A measure of guaranteed availability and its numerical evaluation. IEEE Transactions on Computers 37(1), 25–32 (1988)
6. Huzurbazar, S., Huzurbazar, A.V.: Survival and hazard functions for progressive diseases using saddlepoint approximations. Biometrics 55(1), 198–203 (1999)
7. Lomnicki, Z.A.: A note on the weibull renewal process. Biometrika 53(3/4), 375–381 (1966)
8. Markopoulou, A., Iannaccone, G., Bhattacharyya, S., Chuah, C.N., Ganjali, Y., Diot, C.: Characterization of failures in an operational IP backbone network. IEEE/ACM Transactions on Networking 16(4), 749–762 (2008), doi:10.1109/TNET.2007.902727
9. Muth, E.J.: A method for predicting system downtime. IEEE Transactions on Reliability R-17(2), 97–102 (1968), doi:10.1109/TR.1968.5217522
10. Mykkeltveit, A., Helvik, B.: Adaptive management of connections to meet availability guarantees in SLAs. In: Proceedings of the IM 2009 Mini-Conference (June 2009)
11. Takács, L.: On certain sojourn time problems in the theory of stochastic processes. Acta Mathematica Hungarica 8, 169–191 (1957)
12. Winkelmann, R.: Duration dependence and dispersion in count-data models. Journal of Business & Economic Statistics 13(4), 467–474 (1995)

Application of Intervention Analysis on Stock Market Forecasting

Mahesh S. Khadka[1], K.M. George[2], N. Park[3], and J.B. Kim[4]

[1] Computer Science Department, Oklahoma State University, Stillwater, OK 74078, USA
mahessk@cs.okstate.edu
[2] Computer Science Department, Oklahoma State University, Stillwater, OK 74078, USA
kmg@cs.okstate.edu
[3] Computer Science Department, Oklahoma State University, Stillwater, OK 74078, USA
kmg@cs.okstate.edu
[4] Department of Economics and Legal Studies in Business, Oklahoma State University, Stillwater, OK 74078, USA
jb.kim@okstate.edu

Abstract. In today's financial market, different financial events have direct impact on stock market values. Even a slight change in those events may result a huge difference in stock prices. So consideration of those effects is very important in forecasting stock values. Most of the researches as of now only consider about forecasting but not these effects. This paper studies the effects of some of those events in financial market forecasting. In this paper, we focused our study on the effects of financial events such as GDP, Consumer Sentiments and Jobless Claims on stock market forecasting and analyze them. These events are considered as intervention effects. The intervention effect is described in this study as temporary but immediate and abrupt. So we have tried to estimate not only the period of effect of these events but also use intervening values on forecasting. These forecasted values are then compared to forecasted values obtained from fusion model based on Concordance and Genetic Algorithm (GA). The concept is validated using financial time series data (S&P 500 Index and NASDAQ) as the sample data sets. We also have analyzed how often our forecasting values have the same movement as that of actual market values. The developed tool can be used not only for forecasting but also for in depth analysis of the stock market.

1 Introduction

Stock Market forecasting is considered as one of the most challenging tasks in present financial world. So a lot of attention has been made to analyze and forecast future values and behavior of financial time series. Many factors interact in the stock market including business cycles, interest rates, monitory policies, general economic conditions, traders' expectations, political events, etc. According to academic investigations, movements in market prices are not random. Rather they behave in a highly non-linear, dynamic manner [9]. The ability to predict the direction and values of future stock market is the most important factor in financial market in terms of investment.

Financial events have great impact on day to day behavior of market. So, forecasting models that do not use those effects may not be accurate. GDP, Consumer Sentiments

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 891–899.
springerlink.com

and Jobless Claims etc., are some examples of financial events. Over the years, a lot of researches are being carried out on predicting the stock market. These researches primarily focus on only to develop a new mechanism of forecasting but do not consider the effects of different financial events on forecasting. So in this study, we have proposed a new mechanism that takes care of this issue. Here we have used a forecasting fusion model based on concordance and genetic algorithm (GA) [13] to predict time series in short term in the same or another time series and then analyze and implement effects of financial events that have direct effect on financial market. Whenever these events occur, they immediately affect the market. But these effects do not last forever and for how long they last is unknown. In this study, we have also tried to estimate the duration of their effect on forecasting so that the forecasted values will be as accurate as possible. These intervening effects are time series and can be modeled such that their effects can be applied to market forecasting for better prediction. Based on what are the events that are taken into consideration and their publication date, either a single effect or the joint effects can be used in forecasting. Higher the number of effects taken into consideration, better the forecasting will be.

These days because of online trading, stock market has become one of the hot targets where anyone can earn profits. So forecasting the correct value and behavior of stock market has become the area of interest. However, because of high volatility of the underlying laws behind the financial time series, it is not any easy task to build such a forecasting model [10]. Numbers of forecasting techniques have been proposed so far with their own merits and limitations. Especially the conventional statistical techniques are constrained with the underlying seasonality, non-stationary and other factors [10]. In today's market, people even not of financial sector want to invest their money in stock market and they do not care about how the market is behaving. The only concern they have is whether the value is going up or down so that they can trade their stocks and earn some profit. So, we also have analyzed how often our forecasting values have the same movement as that of actual market values.

2 Previous Approaches to Forecasting

There have been many ways in which the prediction of time series has been proposed, such as extrapolation, linear prediction etc. Some existing models are ARIMA, Box-Jenkins Model, and ARMA etc. Generally there exist two classes of methods of prediction; Parametric Methods and Non-Parametric Methods [1]. The time series data may be stationary or non-stationary as well as seasonal or non-seasonal.

2.1 Parametric Approach

The parametric approach assumes that we can predict the outcome of a time series data based on certain parameters on which the time series is dependent upon. The first stage of such approach typically involves the identification of the parameters on which the data depends. Then, a function or a set of functions on these parameters are constructed. The measures of the parameters are collected from the data and are then used in the set of functions to predict the value of the series.

The parametric approaches are classified into two types based on the types of functions that are used for prediction. They are linear parametric approach and non-linear parametric approach. Linear parametric approach emphasizes that the function or the set of functions defined on the parameters be linear whereas the non-linear parametric approach emphasizes that these functions be non-linear.

Various other approaches are also taken for prediction of time series in economics such as Auto Regressive Moving Average, ARMA, Auto Regressive Integrated Moving Average ARIMA and the Seasonal ARIMA [1]. The ARMA method involves two parts, Auto Regression and the Moving Average, that is, it takes into consideration the regression models of data and also the moving average for analyzing the time series data. The ARIMA method is a generalization of the ARMA model and is obtained by integrating the ARMA model. The data series for ARIMA should be stationary, means it should have constant mean, variance and autocorrelation through time. So series first needs to be differenced until it becomes stationary.

2.2 Non-Parametric Approach

In the Non-Parametric approach, we assume that the data is independent of any other parameters. Some of the Non-Parametric methods that are in use are Multivariate Local Polynomial Regression, Functional Coefficient Autoregressive Model, Adaptive Functional Coefficient Autoregressive Model and the Additive Autoregressive Model [2]. Since the behavior of the varieties decays exponentially with increase in the amount of past data, one of the proposed ways is to convert a multi-dimensional problem into one-dimensional problem by incorporating a single trajectory in the model [8].

Another non parametric approach is the use of perceptron or neural networks [4]. There are many ways to implement such approach. The predictive perceptron model or neural network is created and the historical data is fed as input to the neural network for training. Once the neural network completes the training stage, it can then be used for prediction. Several methods include, conversion of input data into a symbolic representation with grammatical inference in recurrent neural networks to aid the extraction of knowledge from the network in the form of a deterministic finite state automaton [5], preprocessing of input data into Embedded Phase-Space Vectors using delay co-ordinates [6], using special types of networks called Dynamic System Imitator which have been proved to model dynamic complex data. Another method of prediction involves choosing of the training dataset that closely resembles the time series in the ”Correlation Dimension”. In some cases, there are separate neural nets that are used to find undetected regularities in the input dataset [4]. Another way of prediction is to apply a neural network to fuzzy time series prediction using bivariate models to improve forecasting [7].

The advantage of such a system over the parametric approach is that it is very robust, as it can adapt and respond to structural changes. The disadvantage of such an approach is that it can be very data intensive to get fully trained and cannot be used for any data set that is not huge [3].

2.3 Concordance and GA Based Fusion Model

Concordance is defined as the measure of agreement among raters. Given the rating/ranking $X < x_1, x_2, ... >$ and $Y < y_1, y_2, ... >$ given by two judges say, then two pairs of rankings (x_i, y_i) and (x_j, y_j) are said to be concordant if $(x_i, y_i)(x_j, y_j) > 0$. This model is developed in [13]. In this study, we have used Kendall's Tau, Spearman Rho and Gini's Mean Difference [14] for this purpose.

Past data is huge, and needs to be limited to compare with the present using mathematical concordance. The weak Tau, Gini, and Rho concordances of all the possible past segments are compared over a short period of time. This will come out with all the lengths and positions for high concordances. Higher the concordances and longer the matches, indicate better matches. A high concordance means that the trend is likely to continue, so past data can be used to predict the future. To make the prediction as accurate as possible, mathematical equation *g(x)* is searched to map the past data to the future data and to select which section of the past to use based on the concordances. The genetic program will then search for an equation such that $\forall k, g(p_k) \approx f_{k+n}$ where k is a day in the past and n being the offset, in days. Specifically, $\sum(\mathrm{g}(\mathrm{p}_k) - \mathrm{f}_{k+n})^2$ needs to be minimized for all k by choosing the best possible function *g(x)*. The square makes larger differences matter much more than smaller differences. The function *g(x)* will get us close, but it will not be perfect. So the error e_k is measured for each term and subtract that error to get a perfect function. By extrapolating that error and using known values from the past, values that have not happened yet can be guessed. This is done through genetic program.

3 Methodology

The daily changes for market are well fitted by non-Gaussian stable probability density, which is essentially symmetric with location parameter zero. The time evolution of the standard deviation of the daily change of stock market follows power law [11]. The Box-Jenkins model requires data to be stationary. Then seasonality has to be checked. Once stationary and seasonality is addressed, then only identification of order of the autoregressive and moving average terms takes place. The correlation immune to whether the biased or unbiased versions for estimation of the variance are used, concordance is not.

In this section, we discuss about the forecasting methodology. First of all, intervention analysis is done to find out intervention values that are used in forecasting. The comparison between the forecasted values from fusion model and intervention analysis is carried out and find out which one performs better most of the time. Also, the movement of the market is also compared to the movement of forecasted values so that what percentage of time the forecasted values follow the actual movement of the market.

3.1 Intervention Analysis

Intervention analysis study is used to access the impact of a special event on the time series of interest. A simple way to study intervention analysis is to consider some simple dynamic models. To this end, we consider two types of input series. They are (a) the

pulse function and (b) the step function. A pulse function indicates that the intervention only occurs in the single time index t_0 whereas a step function shows that the intervention continues to exist starting with the time index t_0. Consider the weekly sales of a product. Mathematically, these two input functions are

$$P_t^{(t_0)} = \begin{cases} 0; if\, t \neq t_0 \\ 1; if\, t = t_0 \end{cases}$$

$$S_t^{(t_0)} = \begin{cases} 0, & \text{if } t < t_0 \\ 1, & \text{if } t \geq t_0 \end{cases}$$

With a give input, the effect of the intervention can be summarized as

$$f_t = \frac{\omega(B)}{\delta(B)} I_t^{(t_0)}$$

where $I_t = P_t^{(t_0)}$ or $S_t^{(t_0)}$ and $\omega(B) = \omega_0 + \omega_1 B + ... + \omega_s B^s$ or $\delta(B) = 1 - \delta_1 B - .... - \delta_r B^r$ with r and s are non-negative integers. All zeros $Y1,....,Y_{t_0-1}$ of are assumed to be on or outside the unit circle and $\omega(B)$ and $\delta(B)$ have no common factors. Here, we consider function to be step function.

The intervention model can then be written as

$$Y_t = Z_t + \frac{\omega(B)}{\delta(B)} I_t^{(t_0)} \tag{1}$$

Where Y_t is the observed series and Z_t is the underlying intervention-free series. Intervention analysis is to specify the model for Z_t and the polynomial $\omega(B)$ and $\delta(B)$ and so that the impact of the intervention can be estimated.

A general procedure for intervention analysis is as follows:

- Specify the model for Z_t using $\{ Y1,....,Yt_0 - 1 \}$ i.e., using data before intervention.
- Use the model built for Z_t to predict Z_t for $x \geq t_0$. Let the prediction be $\hat{Z}_t$ for $x \geq t_0$.
- Examine Y_t - $\hat{Z}_t$ for $x \geq t_0$ to specify $\omega(B)$ and $\delta(B)$.
- Perform a joint estimation using all the data.
- Check the entertained model for model inadequacy.

In applications, multiple interventions may exist. One can analyze the interventions sequentially before perform a final joint estimation. In this study, we have considered three components; GDP, Consumer Sentiments, and Jobless Claims as intervening factors and their effects on foresting of S&P 500 Index values are calculated and analyzed. For each of considered intervening events, we need to estimate both of the parameters $\omega(B)$ and $\delta(B)$ such that a model can be developed with those parameters using equation (1). The value of is $I_t^{(t_0)}$ after intervention occurs. If we consider only one factor, then (1) will be the model used for forecasting. But if multiple events are considered then parameters of each event should be estimated so that joint estimation of intervention model can be done and used in (1).

4 Experimentation and Result

In this section, we first forecast S&P 500 and NASDAQ index values based on the historic values then effects of GDP, Consumer Sentiments and Jobless Claims are computed and resulting forecasted S&P 500 and NASDAQ index values are calculated.

4.1 Test Data

For the efficacy of the proposed method, we have used stock index values for S&P 500 and NASDAQ from yahoo finance [12]. Table 1 shows the information of the training and test datasets.

Table 1. Training and test data information

Stock Name	Training (In sample Data)		Test (Out of sample) Data	
	From	To	From	To
S&P 500	3 January 1950	31 December 2010	3 January 2011	24 June 2011
NASDAQ	5 February 1971	31 December 2010	3 January 2011	24 June 2011

4.2 Experimental Setup

In this study, forecasting is done for almost about six business months (121 sequential data set) period. However, the forecasting is not done for entire period at once. The idea behind this is, forecasting is done for two business weeks at a time and all the data are collected and analyzed for whole period. This will give more accurate forecast compared to the forecast done for entire year because lower the forecasting horizon, better the forecasting is. Here, we have also compared our forecast with standard fusion model [13] and ARIMA model and shown which method performs better.

The values of intervening events are computed based on the model described in section 3.1. For each of considered intervening events, we need to estimate both of the parameters $\omega(B)$ and $\delta(B)$ such that a model can be developed with those parameters using equation (1). If we consider only one factor, then (1) will be the model used for forecasting. But if multiple events are considered then parameters of each event should be estimated so that joint estimation can be done. After the joint estimation is performed, then (1) is used for forecasting. If the intervening data is published today, then its effect on the market starts tomorrow and lasts for some time. But the problem is its unknown for how long that effect lasts. In this study, we have tried to estimate the period for which this effect lasts. Here, since we have considered Jobless Claims as one of the intervening components and its value is published every week, we have considered the effect will be for a business week. So we have calculated Root Mean Squared Error (RMSE) for 1, 2, 3, 4 and 5 days and compared it with that of forecast without intervention effect. The RMSE is given by

$$RMSE = \sqrt{\frac{\sum_{t=1}^{N} Y_t - F_t}{N}} \tag{2}$$

Table 2. RMSE of forecast with intervention effect for a week

Total Days with intervention effect	RMSE
1	12.36
2	13.48
3	15.10
4	15.66
5	18.08

where, Y is the actual value at time t, F is the forecasted value at time t and N is number of data.

From computation, the RMSE of forecast without intervention effect comes out to be 13.99. Table 2 shows the RMSE until two business days after publication of intervening data is less than that of forecast without considering intervention effect. So from above analysis, we concluded that the effect of intervention can be taken into consideration for two business days after the publication of intervention data. On a given day, if only GDP is published, then only its effect is computed using (1). But if other data are published then the joint effects are computed.

4.3 Result

Table 3 shows the performance improvement of intervention based forecasting model. Here, RMSE of three different models is computed and then compared. Based on statistical notion, model with smaller the RMSE is considered as better model. From the table, we can see that GA based forecasting model performs better than standard ARIMA model in both of the indices. But forecast with intervention model performs better than both of the models.

Fig. 1 and Fig. 2 shows the comparison of actual and forecasted index values for S&P 500 index and NASDAQ index respectively. In the figure, blue line represents the actual index values whereas red color represents the forecasted index values. From the figures, it can be seen that most of the time, the forecasted index values follow the trend of actual index values. It means, if actual value rises then forecasted value also rises and vice versa. From the experimental result, we found out that almost 75% of time the movement of forecasted values is same as the movement of actual values.

Table 3. Performance improvement of intervention model

Stock Index	RMSE for 121 sequential data set		
	Forecast with intervention model	GA based model [13]	ARIMA
S&P 500	7.92	9.90	13.34
NASDAQ	14.80	16.22	63.42

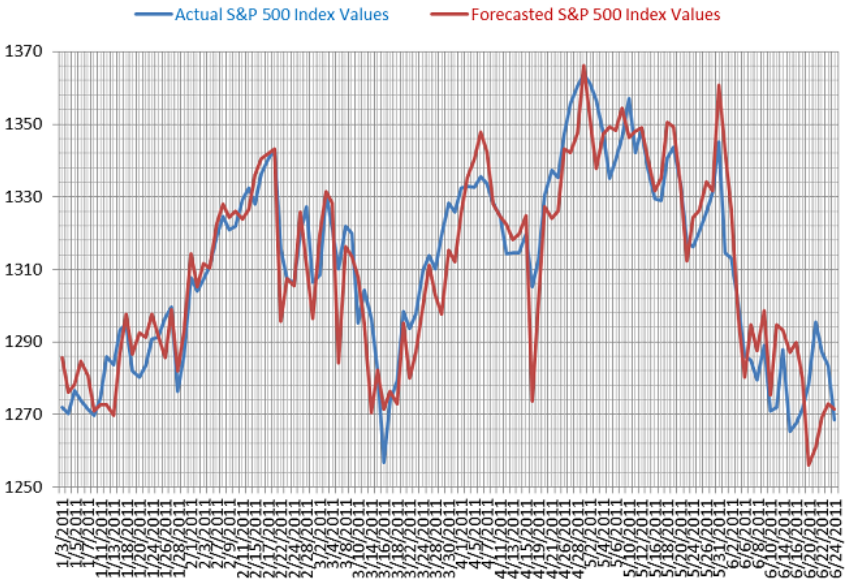


Fig. 1. Comparison of Actual and Forecasted S&P 500 Index Values from 1-3-2011 to 6-24-2011

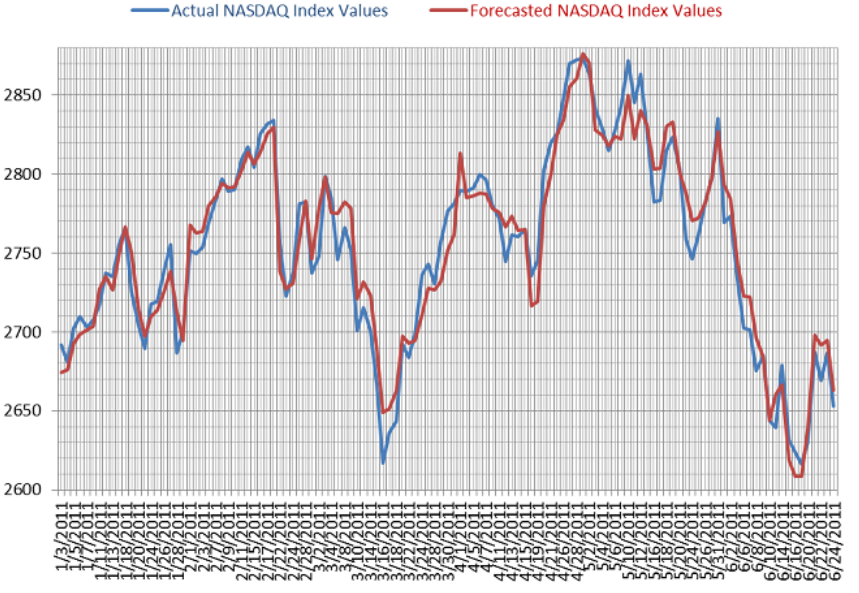


Fig. 2. Comparison of Actual and Forecasred NASDAQ Index Values from 1-3-2011 to 6-24-2011

5 Conclusion

Generally, a comparison between the original time series and a model provides a measure of the model's ability to explain the variability in the original time series. Previous studies tried to optimize the controlling parameters using global search algorithms. Some focus on the optimization of the learning algorithms itself, but most studies had little interest in the elimination of irreverent patterns. Financial events play a huge role in market forecasting. So forecasting model that considers those effect are accurate compared to those that do not. Here in this study, we have come up with a mechanism that takes effects of some of them into consideration such as GDP, Consumer Sentiments and Jobless Claims. From the experimental result, the model with intervention also turns out to be better than the model that does not consider intervention affect. This method performs much better in short forecasting horizon. The case that fusion model performs better is solidified by the conclusion obtained from statistical testing as well.

References

1. Chen, G., Abraham, B., Bennett, G.W.: Parametric and Non-Parametric Modelling of Time Series - An Empirical Study. Environ. Metrics 8, 63–74 (1997)
2. Fan, J., Yao, Q.: Non-Linear Time Series. Springer, New York (2003)
3. Quek, C., Kumar, N., Pasquier, M.: Novel Recurrent Neural Network Based Prediction System for Trading. In: International Joint Conference on Neural Networks, pp. 2090–2097. IEEE (2006)
4. White, H.: Economic Prediction Using Neural Networks: The Case of IBM Daily Stock Returns. In: Proceedings of the 2nd Annual IEEE Conference on Neural Networks, vol. II, pp. 451–458 (1988)
5. Giles, C.L., Lawrence, S., Tsoi, A.C.: Noisy Time Series Prediction using a Recurrent Neural Network and Grammatical Inference. Machine Learning 44, 161–183 (2001)
6. Zhang, J., Chung, S.H., Lo, W.: Chaotic Time Series Prediction Using a Neuro-Fuzzy System with Time-Delay Coordinates. IEEE Transactions on Knowledge and Data Engineering 20, 956–964 (2008)
7. Yu, T.K., Huarng, K.H.: A bivariate fuzzy time series model to forecast the TAIEX. Expert Systems with Applications 34, 2945–2952 (2008)
8. Germán, A., Vieu, P.: Nonparametric time series prediction: A semi-functional partial linear modeling. Journal of Multivariate Analysis 99, 834–857 (2008)
9. Choudhary, R., Garg, K.: A Hybrid Machine Learning System for Stock Market Forecasting. World Academy of Science, Engineering and Technology 39, 315–318 (2008)
10. Hassan, M.R., Kirley, M., Nath, B.: A fusion model of HMM, ANN and GA for Stock Market Forecasting. Expert Systems with Applications 33, 171–180 (2007)
11. Lan, B.L., Tan, O.T.: Statistical Properties of Stock Market Indices of different Economies. Physica A: Statistical Mechanics and its Applications 375, 605–611 (2007)
12. Yahoo Finance Website Historical Prices, `http://finance.yahoo.com/` (cited January 29, 2011)
13. Khadka, M.S., George, K.M., Park, N., Popp, B.: A New Approach for Time Series Forecasting Based on Genetic Algorithm. In: Proceedings of 23rd Annual CAINE Conference, Las Vegas, USA (2010)
14. Kendall, M.: A New Measure of Rank Correlation. Biometrika 30, 81–89 (1938)

Partial Evaluation of Communicating Processes with Temporal Formulas and Its Application

Masaki Murakami

Department of Computer Science,
Graduate School of Natural Science and Technology, Okayama University
3-1-1, Tsushima-Naka, Kita-ku, Okayama, 700-0082, Japan
murakami@momo.cs.okayama-u.ac.jp

Abstract. This paper presents a framework that extends a partial evaluation method for transformational programs to a method for reactive CSP processes. Temporal logic formulas are used to represent constraints on the sets of the sequences of communication actions executed by the processes. We present a set of simple rules for specializing processes with temporal formulas which contain X(next)-operators and/or G(invariant)-operators. We present an example of an application of our partial evaluation method to improve the security of concurrent systems.

1 Introduction

Partial evaluation is a method for optimizing programs using constraints on the input values to improve their efficiency. A number of results are reported in various languages[8]. In the common basic idea of conventional partial evaluation methods, we regard a program as a function from the domain of input values to the range of output values. Let P_f be a program that is an implementation of a function $f(x, y)$. A typical partial evaluation is for example, to obtain a new program P_{fa} from P_f and an input value a, where P_{fa} is the implementation of the function $f(a, y)$. These methods are based on the *transformational* paradigm on which a program is an implementation of a function from the domain of input values to the range of output values.

On the other hand, a program is considered based on the *reactive* paradigm in the area of concurrent computation[5]. A program is not a function from input values to output values on the reactive paradigm. It is a process which communicates with its environment during the computation. For example a process P_1 communicates with the environment by action m_1, then becomes the subsequent process P_2 which makes another action m_2 and so on. In this case, not only m_1 but also m_2 affects to the behaviour of P_2, P_3.. and so on.

Thus we can expect to improve the efficiency of this program by partial evaluation not only with the constraint on the first action m_1 but also with the constraints on the values and the order of actions $m_2, m_3, \ldots$. A number of partial evaluation method for concurrent programs are reported [2,7,10,11,12,16,17]. However, these results do not consider the specialization of concurrent processes wrt the constraints for reactive environments of processes.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 901–909.
springerlink.com

The purpose of this paper is to present a framework that extends a partial evaluation method for transformational programs to a method for reactive processes. We present a simple set of the rules for partial evaluation of reactive CSP[6,15] processes. We use temporal logic formulas[1] for the constraints on the environment of processes. Temporal formulas represent the constraints on the set of sequences of the possible communication actions of processes. The class of formulas used here is the set of temporal formulas which contain X(next)-operators and/or G(invariant)-operators.

Our method makes possible to specialize a concurrent program with the specification of its environment to improve its efficiency. Furthermore, we present an example of an application of our partial evaluation method to improve the security of concurrent systems.

2 Reactive Process and Its Environment

2.1 Reactive Process

We adopt CSP (Communicating Sequential Processes) [6,15] notation to describe reactive processes in this paper.

A process is a function from an action to a process in the framework of CSP. An action is an input communication $c?x$ or an output communication $c!v$, where c is a channel name and x is a variable and v is a term. When $c?x$ in process P and $c!v$ in process Q are executed simultaneously, the value of x becomes v. Thus v is transferred from Q to P using the channel c. A function which maps a communication action m_i to a process P_i is denoted by guarded command like notation such as: $(m_1 \to P_1 [\!] \cdots [\!] m_i \to P_i [\!] \cdots)$ or $[\!]_{i\in I}(m_i \to P_i)$ for a set of indexes I in short[1]. The semantics of a process is defined with a failure set [6,15].

Definition 1 (traces). traces(P) is the set of sequences of actions that a process P can perform defined as follows.

$$\text{traces}([\!]_{i\in I}(m_i \to P_i)) = \{\langle\,\rangle\} \cup \{\langle m_i\rangle^\frown \overline{m} | \overline{m} \in \text{traces}(P_i), i \in I\}$$

$P/\overline{m}$ denotes the process which is obtained from P after execution of $\overline{m}$, namely the continuation of P by $\overline{m}$. $\overline{m_1}^\frown\overline{m_2}$ is the concatenation of a finite sequence $\overline{m_1}$ and a (finite or infinite) sequence $\overline{m_2}$. Namely, if $\overline{m_1} = \langle m_1^1, m_1^2, \ldots, m_1^k\rangle, \overline{m_2} = \langle m_2^1, m_2^2, \ldots, m_2^h\rangle$ (or $\overline{m_2} = \langle m_2^1, m_2^2, \ldots\rangle$) then

$$\overline{m_1}^\frown\overline{m_2} = \langle m_1^1, m_1^2, \ldots, m_1^k, m_2^1, m_2^2, \ldots, m_2^h\rangle$$
$$(\text{or} = \langle m_1^1, m_1^2, \ldots, m_1^k, m_2^1, m_2^2, \ldots\rangle.)$$

$\langle\,\rangle$ is the empty sequence.

Definition 2 (failures). If P is $[\!]_{i\in I}(m_i \to P_i)$ then:

$$\text{failures}(P) = \{(\langle m_i\rangle^\frown\overline{m}, R) | (\overline{m}, R) \in \text{failures}(P_i), i \in I\} \cup$$
$$\{(\langle\,\rangle, R) | \forall j \in I, m_j \notin R\}$$

[1] We consider, for example, $(c?x \to P(x))$ as $[\!]_{i\in I}(c?t_i \to P(t_i))$ where the domain of x is $\{t_i | i \in I\}$ as usual.

If P is an instance $A\theta$ of A by a substitution θ where $A \stackrel{\text{def}}{=} F$, then $\text{failures}(P) = \text{failures}(F\theta)$.

A process which behaves like P_1 or P_2 nondeterministically is denoted as $P_1 \sqcap P_2$. It is impossible from the environment to control which of P_1 or P_2 is selected.

Definition 3 (nondeterministic process). $\text{failures}(P_1 \sqcap P_2) = \text{failures}(P_1) \cup \text{failures}(P_2)$.

2.2 Environments of Reactive Processes

Let P and P' be processes. Consider that there is no difference between the sets of possible behaviors of processes P and P' for a given environment. In other words, P and P' cannot be distinguished under the environment. If P' can be executed more efficiently than P, then we should adopt P' rather than P when it is executed in the environment.

The environment of a reactive process P is the set of processes which communicate with P. The constraints on the environments are derived from the specification of the environment processes that communicate with the target process. The specification of environment processes can be given with temporal logic formulas[1]. We use temporal operators such as the "always"-operator G and/or the "next"-operator X in formulas[2].

The truth value of a temporal formula is defined on a sequence which denotes the time axis. Usually, the truth values of formulas without temporal operators that represent the constraints on the input value are defined using the notion of state of a program which is decided from the values of input variables. However, truth values of formulas without temporal formulas are not essential here, we avoid to discuss that in detail. We assume that for any formula p without temporal operators, the set *modelof*(p) of sequences that make p true on them is defined.

In this paper, a time axis is a trace of a process. Let $\overline{m}$ be a finite sequence of communication actions $\langle m_1, m_2, \ldots, m_n\rangle$ or an infinite sequence $\langle m_1, m_2, \ldots, m_i, \ldots\rangle$. We denote $\mathit{head}(\overline{m}) = m_1$, and $\mathit{tail}_i(\overline{m}) = \langle m_i, \ldots, m_n\rangle$(or $\langle m_i, \ldots\rangle$). Note that $\mathit{tail}_1(\overline{m}) = \overline{m}$.

Definition 4. Let $\overline{m}$ be a sequence of communication actions.

1. If $p(x)$ is a predicate formula which does not contain temporal operators, and x is a variable on the set of communication actions or a variable that occurs in communication actions, $\overline{m} \models p$ iff $\overline{m} \in \text{modelof}(p)$.
2. $\overline{m} \models \mathsf{X}p$ iff $\text{tail}_2(\overline{m}) \models p$
3. $\overline{m} \models \mathsf{G}p$ iff $\forall i(1 \leq i), \text{tail}_i(\overline{m}) \models p$

We introduce the predicate *done*(m). Intuitively, if *done*(m) is true on a state then the action m is "just finished" before reaching the state. For a substitution θ and a term v, we denote $c!(v\theta)$ as $m\theta$ if $m = c!v$. Similarly, we denote $c?(x\theta)$ as $m\theta$ if $m = c?x$ for a variable x.

[2] We do not consider formulas that contain F (possible) operators.

Definition 5. $\langle m_0 \rangle^\wedge \overline{m'} \models \mathsf{X}\mathrm{done}(m)$ iff $m\theta = m_0$ for some substitution θ.

Definition 6. Let M be a set of sequences of communication actions.

$$M \models p \text{ iff } \forall \overline{m} \in M, \overline{m} \models p$$

For a process P and a formula p, we denote $P \models p$ iff $M \models p$ where M is the set of sequences of communication actions such that P executes with the environment. Namely

$$P \models p \text{ iff } \forall t \in \mathrm{traces}(P), t \models p.$$

3 Set of Rules for Partial Evaluation

Let P be a process and p be a temporal formula which represents a constraint for the environment of P. We denote the partial evaluation of process P by p as *Part*(P, p). This section presents a method to obtain a new process which is equal to *Part*(P, p).

Let P' be a new process name and let P' be: $P' \stackrel{\text{def}}{=} \mathit{Part}(P, p)$. We rewrite the right hand side of this definition using the following rules.

- **Unfolding / Folding**
  Unfolding $P\theta \Rightarrow F\theta$
  Folding $F\theta \Rightarrow P\theta$
  where $P \stackrel{\text{def}}{=} F$ and $P\theta$ (or $F\theta$) means an instance of P (or F respectively) which is obtained by applying a substitution θ.
- **Non-temporal Logical Rules**
  Predicate rule $\mathit{Part}(P, p) \Rightarrow P_p$
  where p is a predicate formula without temporal operators, and P_p is obtained with partial evaluation from P and p (by a conventional method).
  ***done*-rule** $(m \to P) \Rightarrow (m \to \mathit{Part}(P, \mathit{done}(m)))$
  $\wedge^-$**-rule** $\mathit{Part}(P, p \wedge q) \Rightarrow \mathit{Part}(\mathit{Part}(P, p), q)$
  $\wedge^+$**-rule** $\mathit{Part}(\mathit{Part}(P, p), q) \Rightarrow \mathit{Part}(P, p \wedge q)$
  $\supset$**-rule** $\mathit{Part}(P, p) \Rightarrow \mathit{Part}(P, q)$ if $p \supset q$
- **Temporal Rules**
  Let P be a process which is equal to the function $[\!]_{i \in I}(m_i \to P_i)$
  X-rule $\mathit{Part}(P, \mathsf{X}p) \Rightarrow [\!]_{i \in I}(m_i \to \mathit{Part}(P_i, p))$
  G-rule $\mathit{Part}(P, \mathsf{G}p) \Rightarrow \mathit{Part}([\!]_{i \in I}(m_i \to \mathit{Part}(P_i, \mathsf{G}p)), p)$
  Pruning rule $\mathit{Part}(P, \mathsf{X}\mathit{done}(m_k\theta)) \Rightarrow (m_k\theta \to P_k)$
  for $k \in I$ if $m_j\theta' \neq m_k\theta$ for any substitution θ' and any $j(\in I) \neq k$.
- $\sqcap$**-rule**

$$\mathit{Part}(P_1 \sqcap P_2, p) \Rightarrow \mathit{Part}(P_1, p) \sqcap \mathit{Part}(P_2, p)$$

- **Termination rule**

$$\mathit{Part}(P, p) \Rightarrow P$$

4 Soundness

In this section, we prove that the process obtained by the transformation in the previous section behaves similarly to the original process under the constraint.

4.1 Restricted Failure Set Equivalence

We introduce *restricted failure set equivalence* to formalize the notion that two processes behave equivalently under a given constraint. We also show a number of properties of the equivalence. Restricted failure set equivalence is defined using the notion of failure set equivalence.

Definition 7 (Restriction of a failure set). Let F be a set of pairs $(\overline{m}, R)$ where $\overline{m}$ is a sequence of actions and R is a set of actions, and let q be a temporal formula. $F \downarrow q$ is a *restriction of F by q* defined as follows.

$$F \downarrow q = \{(\overline{m}, R) | \exists r, (\overline{m}, R) \in F, \mathrm{head}(r) \in R, \overline{m}^{\wedge} r \models q\}$$

Definition 8 (Restricted Failure Set Equivalence). Let q be a temporal formula. Processes P_1 and P_2 are *restricted failure set equivalent wrt q* if:

$$\mathrm{failures}(P_1) \downarrow q = \mathrm{failures}(P_2) \downarrow q$$

and denoted $P_1 \sim P_2$ wrt q.

If $P_1 \sim P_2$ wrt q then they behave similarly and no deference can be observed under the environment which satisfies q. The following propositions are easy to prove.

Proposition 1. *$\cdot \sim \cdot$ wrt q is an equivalence relation for any q.*

Proposition 2. *1. Let $P \stackrel{def}{=} F$.*
i) If $F\theta \sim P'$ wrt q then $P\theta \sim P'$ wrt q.
ii) If $P\theta \sim P'$ wrt q then $F\theta \sim P'$ wrt q.
2. Let P be a process that is equal to the function $[\!]_{i \in I}(m_i \to P_i)$. If $P_i \sim P_i'$ wrt q for every $i \in I$, then $P \sim [\!]_{i \in I}(m_i \to P_i')$ wrt $\mathsf{X}q$.
3. If $P \sim P'$ wrt q, then $P \sqcap Q \sim P' \sqcap Q$ wrt q and $Q \sqcap P \sim Q \sqcap P'$ wrt q

These results show that restricted failure set equivalence is a congruence relation for above operations.

Proposition 3. *1. If $P \sim P'$ wrt $done(m)$, then $failures((m \to P)) = failures((m \to P'))$.*
2. If $P \sim P'$ wrt q_1 and $P' \sim P''$ wrt q_2 , then $P \sim P''$ wrt $q_1 \wedge q_2$.
3. For any P and P', if $P \sim P'$ wrt $p \wedge q$, then there exists P'' such that $P \sim P''$ wrt p and $P'' \sim P'$ wrt q.
4. If $P \sim P'$ wrt q and $p \supset q$, then, $P \sim P'$ wrt p.
5. $[\!]_{i \in I}(m_i \to P_i) \sim (m_k\theta \to P_k)$ wrt $\mathsf{X}done(m_k\theta)$ for $k \in I$ if $m_j\theta' \neq m_k\theta$ for any substitution θ' and any $j(\in I) \neq k$.

4.2 Soundness of Transformation Rules

We assume that the soundness of the conventional partial evaluation method is already shown. In other words, we assume that if P' is the result of partial evaluation of P by (non-temporal) constraint q, then $P \sim P'$ wrt q.

Let $E_0 = \mathit{Part}(P, p)$. A finite sequence $E_0 \Rightarrow E_1 \Rightarrow E_2 \Rightarrow \ldots \Rightarrow E_n$ is a *transformation sequence* if E_i is obtained from E_{i-1} immediately by applying one of the rules in section *3.1*. If E_n no longer contains a sub expression in the form of $\mathit{Part}(Q, q)$, then the partial evaluation is completed.

We can show the following proposition by the induction on n using **Proposition 2** and **3**.

Proposition 4. *Let E_0 be $\mathit{Part}(P, p)$. For a transformation sequence $E_0 \Rightarrow E_1 \Rightarrow E_2 \Rightarrow \ldots \Rightarrow E_n$, let P' be a process which is obtained by replacing all sub-expressions $\mathit{Part}(Q_1, q_1), \ldots, \mathit{Part}(Q_k, q_k)$ of the form $\mathit{Part}(...)$ in E_n with $Q'_1, \ldots, Q'_k$ respectively. If $Q_i \sim Q'_i$ wrt q_i $(1 \leq i \leq k)$, then $P \sim P'$ wrt p.*

Theorem 1 (The soundness). *For any process P and the constraint p, if Q is obtained from $\mathit{Part}(P, p)$ by applying the rules, then $P \sim Q$ wrt p.*

Proof. Q is E_n of **Proposition 4** without sub-expressions of the form $\mathit{Part}(\cdots)$.

5 Improvement of Security

This section presents an example of an application of our partial evaluation method to improve the security of a concurrent system. It is an example of authentication protocol that is a modification of Needham-Schroeder-Lowe protocol[4,9].

Example 1. Consider an example consists of three agents *Alice, Bob* and *Charlie*. Alice and Bob establish a connection with authentication using public key cryptosystem, and let Charlie be *a man in the middle*. Every communication between Alice and Bob is transferred via Charlie. First, we consider the case that Charlie is not malicious and he just forwards massages from Alice to Bob or from Bob to Alice. In this case, Charlie is regarded as a part of the trusted network.

The protocol is as follows. First, Alice sends a request R to get Bob's public key K_B. Bob sends K_B as the reply to R. Alice sends the message $[\mathit{Id}_A, N_A]_{K_B}$ encrypted with K_B where Id_A is her own id, N_A is a nonce. Bob receives $[\mathit{Id}_A, N_A]_{K_B}$ and decrypt the message. Then he gets N_A. He sends the message $[\mathit{Id}_B, N_A, N_B]_{K_A}$ consists of his own id Id_B, N_A and a new nonce N_B encrypting with Alice's public key K_A (We assume Bob already has Alice's public key for the simplicity). Alice receives the message and she gets N_B. She sends N_B encrypting with K_B as $[N_B]_{K_B}$. Then Bob receive $[N_B]_{K_B}$ and then the authentication is completed.

The system consists of Alice and Bob is defined as the process *AliceBob* that communicates with the process of Charlie. Let c_A be the channel name for communication from Alice to Charlie, and a be the the channel from Charlie to Alice. c_B and b

are channels for the communication of Charlie and Bob. In the followings, terms begin with capital letters such as $N_A, N_B, Id_A, Id_A, K_A, K_B$ and K_C denotes values of nonces, id's or public keys respectively. Terms begin with lower case letters such as $id_A, id_B, n_A, n_B, k, k_B, \ldots$ are variables to receive id's, nonces or keys respectively.

$$\begin{aligned} \mathit{AliceBob} \stackrel{\text{def}}{=} & (c_A!R \to (b?r \to (c_B!K_B \to (a?k \to \\ & (c_A![Id_A, N_A]_k \to (b?[id_A, n_A]_{K_B} \to \\ & (c_B![Id_B, n_A, N_B]_{K_A} \to (a?[id_B, N_A, n_B]_{K_A} \to \\ & (c_A![n_B]_k \to (b?[N_B]_{K_B} \to \mathrm{OK})))))))))) \end{aligned}$$

In this case, Alice is so incautious that she ignores the value of id_B and does not check the validity of public key k just as the original Needham-Schroeder protocol[14] with the security hole which [9] reported.

The man in the middle Charlie is the process as follow if he is not malicious.

$$\begin{aligned} C \stackrel{\text{def}}{=} & (c_A?r \to (b!r \to (c_B?k_B \to (a!k_B \to (c_A?m_A^1 \to (b!m_A^1 \to \\ & (c_B?m_B \to (a!m_B \to (c_A?m_A^2 \to b!m_A^2 \to C))))))))) \end{aligned}$$

where r is a variable for the request, $m_A^j (j = 1, 2)$ are variables for the messages from Alice and m_B is a variable for the message from Bob. He just forwards the messages for each direction.

On the other hand, we consider the case that Charlie is malicious. The process $C' \stackrel{\text{def}}{=} C \sqcap C''$ is the behaviour of malicious Charlie making the "man-in-the-middle attack" where

$$\begin{aligned} C'' \stackrel{\text{def}}{=} & (c_A?r \to (b!r \to (c_B?k_B \to (a!K_C \to (c_A?m_A^1 \to (b![a, n_A]_{K_B} \to \\ & (c_B?m_B \to (a!m_B \to (c_A?m_A^2 \to (b![n_B]_{K_B} \to \mathit{fakeBob}))))))))))). \end{aligned}$$

As C' is a nondeterministic process, sometimes he may act as C, but he may act maliciously. In the case of malicious Charlie, C' replies his own public key K_C instead of Bob's key for Alice's request. As the Alice's message m_A^1 is $[Id_A, N_A]_{K_C}$ that is encrypted with Charlie's key K_C, he can decrypt it and get A and N_A. He send Bob them as $[a, n_A]_{K_B}$. After forwarding Bob's reply to Alice, he get N_B from m_A^2 as it is encrypted with K_C. So he get N_A and N_B, he can pretend to be Bob.

If Alice is cautious, the system of Alice and Bob is defined as follows.

$$\begin{aligned} \mathit{AliceBob}' \stackrel{\text{def}}{=} & (c_A!R \to (b?r \to (c_B!K_B \to (a?k \to \\ & (c_A![Id_A, N_A]_k \to (b?[id_A, n_A]_{K_B} \to \\ & (c_B![Id_B, n_A, N_B]_{K_A} \to \\ & \quad (a?[\text{“the id of the owner of } k\text{”}, N_A, n_B]_{K_A} \to \\ & (c_A![n_B]_k \to (b?[N_B]_{K_B} \to \mathrm{OK})))))))))) \end{aligned}$$

She checks Bob's message if the id is same to the owner of the received key k. If Charlie is as C (not malicious), k is Bob's key and the id of the owner of k is equal to Id_B that

she receives. If the owner of k is not equal to the Id_B, the she refuses to receive the message and does not proceed the protocol.

For these definitions, when *AliceBob*' behaves satisfying *goodC* such that:

$$\begin{aligned} \mathit{goodC} \equiv\ & \mathsf{X}(\mathit{done}(c_A!R) \wedge \mathsf{X}(\mathit{done}(b?R) \wedge \\ & \mathsf{X}(\mathit{done}(c_B!K_B) \wedge \mathsf{X}(\mathit{done}(a?K_B) \wedge \\ & \mathsf{X}(\mathit{done}(c_A!m_A^1) \wedge \mathsf{X}(\mathit{done}(b?m_A^1) \wedge \\ & \mathsf{X}(\mathit{done}(c_B!m_B) \wedge \mathsf{X}(\mathit{done}(a?[\mathit{Id}_B, N_A, n_B]_{K_A}) \wedge \\ & \mathsf{X}(\mathit{done}(c_a!m_A^2) \wedge \mathsf{X}(\mathit{done}(b?m_A^2)))))))))), \end{aligned}$$

C' acts as C because Alice receives K_B at $a?k$ that is same to the value sent by $c_B!K_B$ and the id of owner of this key is Id_B that she receives. Then Alice and Bob establish the connection and we have *AliceBob* $\sim$ *AliceBob*$'$ wrt *goodC*. Namely, they are equivalent if Charlie is not malicious.

On the other hand, when C' does $a!K_C$ after $c_B?k_B$, *AliceBob*$'\|C'$ behaves differently from *AliceBob*$\|C'$. It deadlocks without establishing the insecure connection.

Thus, we consider that for a process P, the secure version of P is obtained as the process P' such that P' detects attacks and $P' \sim P$ wrt q where q is the constraint that there is no malicious agent who communicates with P or P'. Our partial evaluation method gives a process Q such that $Q \sim P$ wrt q as presented in the previous section. So we consider that the partial evaluation method is useful to improve not only the efficiency of process but also to improve the security of system.

In the case of this example, both of *Part*(*AliceBob*$'$, *goodC*) and *Part*(*AliceBob*, *goodC*) can be transformed into the following process *AliceBob*$''$ using **X-rule** and **Pruning rule** etc. presented in section 3. Note that in this process, Alice knows K_B and Id_B before she receive them.

$$\begin{aligned} \mathit{AliceBob}'' \stackrel{\text{def}}{=} & (c_A!R \to (b?r \to (c_B!K_B \to (a?K_B \to \\ & (c_A![\mathit{Id}_A, N_A]_{K_B} \to (b?[\mathit{id}_A, n_A]_{K_B} \to \\ & (c_B![\mathit{Id}_B, n_A, N_B]_{K_A} \to (a?[\mathit{Id}_B, N_A, n_B]_{K_A} \to \\ & (c_A![n_B]_{K_B} \to (b?[N_B]_{K_B} \to \mathrm{OK}))))))))) \end{aligned}$$

From **Theorem 1**, we have *AliceBob*$'' \sim$ *AliceBob* wrt *goodC* and *AliceBob*$'' \sim$ *AliceBob*$'$ wrt *goodC*. Then we have *AliceBob* $\sim$ *AliceBob*$'$ wrt *goodC* from **Proposition 1**.

6 Conclusion

A set of the rules for partial evaluation of CSP processes using temporal formulas and its soundness are presented. We can specialize reactive processes using constrains on the input messages which are delivered in the middle of execution. The set of rules presented here can be used with any conventional partial evaluation method for CSP like languages if it preserves restricted failure set equivalence. Thus the set of rules

of this paper can be regard as a framework to extend partial evaluation methods for transformational programs to reactive processes. We also mentioned the relation to the improvement of security.

References

1. Emerson, E.A.: Temporal and Modal Logic. In: van Leeuwen, J. (ed.) Handbook of Theoretical Computer Science. Formal Models and Semantics, vol. B, ch. 16, pp. 995–1072. The MIT Press/Elsevier (1990)
2. Etalle, S., Gabbrieli, M.: Partial evaluation of concurrent constraint languages. ACM Computing Surveys 30(3) (September 1998)
3. Glück, R., Jøgensen, J., Martens, B., Sørensen, H.: Controlling Conjunctive Partial Deduction. In: Kuchen, H., Swierstra, S.D. (eds.) PLILP 1996. LNCS, vol. 1140, pp. 152–166. Springer, Heidelberg (1996)
4. Hagiya, M., *Introduction for Cryptosystem* The University of Tokyo (2010) (in Japanase), `http://hagi.is.s.u-tokyo.ac.jp/pub/staff/hagiya/kougiroku/ango/intro.pdf`
5. Harel, D., Pnueli, A.: On the Development of Reactive Systems. In: Apt, K.R. (ed.) Logic and Models of Concurrent Systems. Springer, New York (1985)
6. Hoare, C.A.R.: Communicating Sequential Processes. Prentice Hall (1985)
7. Hosoya, H., Kobayashi, N., Yonezawa, A.: Partial Evaluation Scheme for Concurrent Languages and its Correctness. In: Fraigniaud, P., Mignotte, A., Bougé, L., Robert, Y. (eds.) Euro-Par 1996. LNCS, vol. 1123, pp. 625–632. Springer, Heidelberg (1996)
8. Jones, N.D., Gomard, C.K., Sestoft, P.: Partial Evaluation and Automatic Program Generation. Prentice-Hall (1993)
9. Lowe, G.: An attack on the Needhum-Schroeder public-key authentication protocol. Information Processing Letters 56, 131–133 (1995)
10. Marinescu, M., Goldberg, B.: Partial-evaluation techniques for concurrent programs. In: Proc. of ACM SIGPLAN PEPM 1997, pp. 47–62 (1997)
11. Martel, M., Gengler, M.: Partial Evaluation of Concurrent Programs. In: Sakellariou, R., Keane, J.A., Gurd, J.R., Freeman, L. (eds.) Euro-Par 2001. LNCS, vol. 2150, pp. 504–514. Springer, Heidelberg (2001)
12. Masuhara, H., Yonezawa, A.: Design and Partial Evaluation of Meta-objects for a Concurrent Reflective Language. In: Jul, E. (ed.) ECOOP 1998. LNCS, vol. 1445, pp. 418–439. Springer, Heidelberg (1998)
13. Murakami, M.: Partial Evaluation of Reactive Concurrent Processes using Temporal Logic Formulas. Computer Software 12(3), 15–27 (1995)
14. Needhum, R.M., Shroeder: Using encryption for authentication in large network of computers. Communications of the ACM 21(12), 993–999 (1978)
15. Roscoe, A.W.: Theory and Practice of Concurrency. Prentice-Hall (1998)
16. Sugahara, T., Watanabe, T.: A Method for Partial Evaluation of Object-Oriented Concurrent Languages. In: Proc. 12th Conf. JSSST, pp. 337–340 (1995)
17. Zöbel, D.: Program Transformations for Distributed Control Systems, `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.73.5403`

Performance Analysis of a Hybrid Photovoltaic Thermal Single Pass Air Collector Using ANN

Deepali Kamthania*, Sujata Nayak, and G.N. Tiwari

Center for Energy Studies, Indian Institute of Technology Delhi,
Hauz Khas, New Delhi – 110016, India

Abstract. This paper presents the performance analysis of semi transparent hybrid photovoltaic single pass air collector considering four weather conditions (a, b, c and d type) of New Delhi weather station of India using ANN technique. The MATLAB 7.1 neural networks toolbox has been used for defining and training of ANN for calculations of thermal, electrical, overall thermal energy and overall exergy. The ANN models use ambient temperature, number of clear days, global and diffuse radiation as input parameters. The transfer function, neural network configuration and learning parameters have been selected based on highest convergence during training and testing of networks. About 3000 sets of data from four weather stations (Bangalore, Mumbai, Srinagar, and Jodhpur) have been given as input for training and data of the fifth weather station (New Delhi) has been used for testing purpose. ANN model has been tested with Levenberg-Marquardt training algorithm to select the best training algorithm. The feedforward back-propagation algorithm with logsig transfer function has been used in this analysis. The results of ANN model have been compared with analytical values on the basis of root mean square error.

Keywords: Hybrid photovoltaic thermal (HPVT), Single pass air collector (SPAC), Log Sigmoid (logsig), Root Mean Square Error (RMSE), Lavenberg - Marguardt (LM), Artificial Neural Network (ANN).

1 Introduction

Rapidly depleting rate of conventional energy had led to the exploration of possibility for utilizing non conventional energy sources. Solar energy conversion is one of the most promising renewable energy technologies which have potential to contribute significantly to sustainable energy supply. The major component of any solar energy system is solar air collector, which absorb solar radiation energy and transfer it to transport medium. Solar photovoltaic technology, utilizes solar energy in the form of electrical as well as thermal energy after its conversion. Jiang [1] has developed a model to estimate monthly mean daily diffuse solar radiation for eight typical cities in China. It has been observed that ANN based estimation technique is more suitable than the empirical regression models. Leal et al. [2] has measured, analyzed and compared three different statistical models and two ANN model for estimating the

* Corresponding author.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 911–918.
springerlink.com

daily UV solar radiation from the daily global radiation. It has been observed that the statistical and ANN models have good statistical performance with RMSE lower than 5% and MBE between 0.4 - 2 %. Koca et al. [3] has developed an ANN model for estimation of future data on solar radiation for seven cities from Mediterranean region of Anatolia in Turkey. It has been observed that the results obtained can be used to design high efficiency solar devices. Sencan et al. [4] have developed ANN models for predicting thermal performance of solar air collector. Experimental data has been used for training and testing of the networks with input as inlet air temperature, solar radiation, air mass flow rate and the output as thermal performance of solar air collector. The R2-values have been observed as 0.9985 for unknown data. Caner et al. [5] have designed an ANN model to estimate thermal performances of two types of solar air collectors. The calculated and predicted values of thermal performances have been compared and statistical error analysis has been carried out to demonstrate effectiveness of the proposed ANN model.

In this paper an attempt has been made to develop ANN models to analyze the performance of a semi transparent HPVT-SPAC considering different type of weather conditions. The four type of weather conditions are defined, Singh [6]. The data of solar radiations for different climates obtained from Indian Metrological Department (IMD), Pune for four weather stations (Bangalore, Mumbai, Srinagar, and Jodhpur) have been used for training and data of the fifth weather station (New Delhi) has been used for testing purpose. The results of ANN models for semi transparent HPVT-SPAC have been compared with analytical vales on the basis of RMSE.

2 Description and Design of ANN

The MATLAB 7.1 neural network toolbox has been used to develop an ANN model. In order to train the network the data of solar radiations for different climates has been obtained from IMD, Pune. Fig.1. shows the structure of ANN model. The model uses ambient temperature, number of clear days, global and diffuse radiation for a climatic condition as input parameters and electrical, thermal, overall thermal energy and overall exergy as output parameters for the experimental setup is shown in Fig. 3. The network consist of three layers the input layer, hidden layer and output layer with 12 neurons in the hidden layer and 4 neurons in input and output layer respectively. The first hidden layer has 'tan-sigmoid' activation function Φ defined by the logistic function as $\phi = 1/1 - e^{n}$, where n is the corresponding input. Fig.2. represents the typical layout of the ANN, which shows the network nodes along with biases and weights. The network type is selected as feed forward back propagation. TRAINLM has been selected as training function and MSE has been taken as the performance function. The MSE training goal has been set to 0.005 as shown in Fig. 4. The inputs have been normalized in the (0, 1) range. LOGSIG has been taken as transfer function between layers. A set of 3000 epochs has been taken for training purpose. This training function updates the weights and bias values in accordance with Levenberg–Marquardt optimization. In order to train the network the data of solar radiations for different climates has been obtained from IMD, Pune. The following parameters are set while training the feed forward neural network: training pattern 3000, learning rate 0.001, MSE training goal has been set as 0.005, number of training iterations 125,

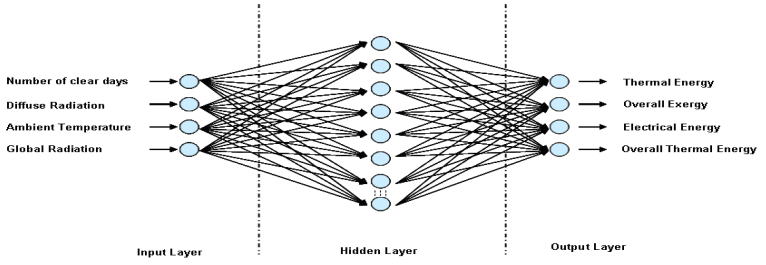


Fig. 1. Input, output and hidden layers of ANN

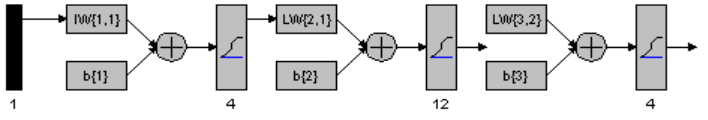


Fig. 2. Typical arrangement of ANN

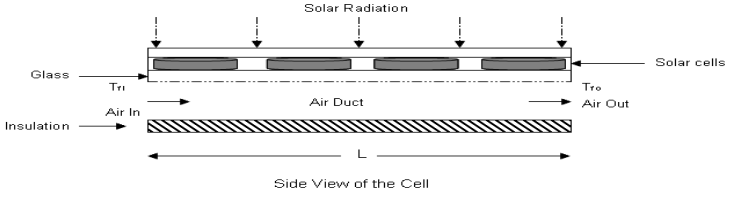


Fig. 3. Schematic diagram of HPVT- SPAC

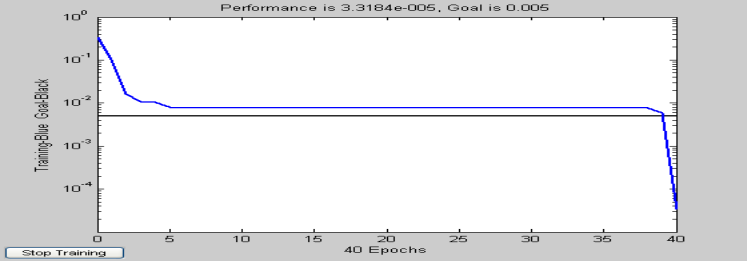


Fig. 4. MSE obtained in the training of the network

momentum 0.91. The training patterns are presented repeatedly to the ANN model and the adjustment is performed after each iteration whenever the network's computed output is different from the desired output. It has been observed that LM with 12 neurons in the hidden layers is the most suitable algorithm for semi transparent HPVT-SPAC.

3 Energy and Exergy Analysis of Single Pass Air Collector

A cross sectional view of semi transparent HPVT-SPAC for the experiential setup installed at the Solar energy park of IIT, New Delhi, illustrated in Fig. 3 has duct placed below the PV module. The air is passed through one end of the air duct collector and gets warm by picking the thermal energy from the back side of the PV module and exit from the other end of the duct. The lower duct is insulated to minimize the heat loss.

The annual thermal energy output can be evaluated as

$$Q_{thannual} = \sum_{k=1}^{12} \frac{\dot{q}_u}{1000} \times N \times n_0 \tag{1}$$

The annual exergy can be calculated as

$$Ex_{thannual} = Q_{thannual} \left[1 - \frac{\overline{T}_a + 273}{\overline{T}_{fo} + 273} \right] \tag{2}$$

The annual electrical energy can be obtained as

$$(E_{el})_{annual} = \eta_{el} \times A \times I(t)_{avg} \times N \times n_0 \tag{3}$$

The overall thermal energy can be expressed as

$$(Q_{ov})_{th} = Q_{thannual} + \frac{(E_{el})_{annual}}{\mathbf{0}.38} \tag{4}$$

The annual exergy can be obtained from Eqs. (2) and (3)

$$Ex_{annual} = Ex_{thannual} + (E_{el})_{annual} \tag{5}$$

For details, please refer to paper written by Kamthania et al. [7, 8].

4 Methodology

The ANN has been defined in MATLAB 7.1 neural network toolbox as per the above mentioned parameters. The initial values of the weights have been defined and an incremental input is given to the network for estimating the outputs. When the outputs are closure to result matrix and the calculated MSE is within specified limits the iterations are terminated and the values of weights are recorded. When the output matrix is close to desired results then the network is trained otherwise the same procedure is to be repeated with new weight matrix. Thus value obtained through ANN model are compared with analytical result for New Delhi weather station. The RMSE deviation has been calculated using the following equation.

$$e = \left(\sqrt{\frac{\sum e_i^2}{n}} \right) \times 100 \qquad \text{where} \quad e_i = \left[\frac{X_i - Y_i}{X_i} \right] \tag{6}$$

5 Results and Discussion

The purpose of this study is to develop an ANN model for performance analysis and compare the results with analytical study for semi transparent HPVT-SPAC. Fig. 4 shows MSE curve for a typical iteration, the performance of the network has been shown against the goal set for the network. MSE has been taken as the performance function with MSE training goal set as 0.005. It has been observed that LM with 12 neurons in the hidden layer for HPVT-SPAC and 4 neurons in input and output layer is the most suitable algorithm with set MSE value. The RMSE measures the average magnitude of error. It is better to have lower RMSE values. The RMSE has been calculated using Eq.6. The RSME values of the performance parameters calculated from both ANN model and analytical study for different months considering a, b c and d type weather conditions have been shown in Table 1. According to the results the deviation are in the range of 0.10-2.23% for different output parameters. It has been observed that from Table 2 that RMSE for thermal energy (Q_u), electrical energy (E_{el}), overall thermal energy (Q_{ovth}) and overall exergy ($Ex_{monthly}$) varies from 0.122 to 1.822%, 0.52 to 1.66%, 1.03to 1.60 %0.19 to 1.47% for SPAC respectively. Table 3 shows the number of clear days in different weather condition for New Delhi weather station. It has been observed that that maximum value of thermal energy, electrical energy, overall thermal energy and overall exergy are obtained for 'c' type weather condition due to maximum number of clear days whereas minimum value of thermal energy, electrical energy, overall thermal energy and overall exergy have been obtained for'd' type weather condition due to minimum number of clear days. Fig. 5, 6 and 7 shows deviations of various performance parameters. Fig. 5 shows the monthly variation of thermal energy for a, b, c and d type weather conditions for New Delhi for SPAC. It has been observed that for SPAC analytical value is 293.44 kWh whereas ANN values are 288.72kWh for c type weather condition. Fig. 6 shows the monthly variation of thermal and electrical energy for semi transparent HPVT-SPAC. It has been observed that thermal and electrical energy is 827.66 kWh and 797.99 kWh respectively for semi transparent HPVT-SPAC. Fig. 7 shows the analytical values of annual overall thermal energy and exergy for New Delhi. The analytical values of annual overall thermal energy and exergy for semi transparent HPVT-SPAC are 2915. 52 and 840.68 kWh respectively whereas 2896.27 and 823.27 kWh respectively with ANN method for New Delhi. The ANN values obtained in Fig. 5, 6 and 7 are also very close to the analytical values. It has been observed that ANN simulation values can be obtained easily without involving complex computations.

Table 1. RMSE calculations for different months considering a, b c and d type weather conditions for New Delhi

Months	Error ∑Qua SP, kWh	Error ∑Qub SP, kWh	Error ∑Quc SP, kWh	Error ∑Qud SP, kWh
Jan	0.0210	-0.1990	-0.1000	-0.0996
Feb	-0.0543	-0.0081	1.0000	0.1909
March	0.2000	0.0523	0.2626	-0.3757
April	1.0000	-0.9920	2.2000	-0.4040
May	0.0000	-2.0000	-0.0700	-1.1025
June	-0.1000	2.0000	0.1900	1.1584
July	-0.1286	-1.0200	-1.0000	1.7337
Aug	0.0315	-1.0000	-1.0298	-0.6705
Sep	-0.2970	1.0000	-1.0585	-0.5280
Oct	-0.2300	3.0000	2.0000	0.4892
Nov	0.1600	0.8909	1.3248	1.3123
Dec	0.1014	-0.1200	1.0000	2.7584

Table 2. RMSE calculations for different parameters considering a, b c and d type weather conditions for New Delhi

Months	$\sum Q_{u\ SP}$, kWh	$\sum E_{el\ SP}$, kWh	$Q_{ovth\ SP}$, kWh	$Ex_{monthly\ SP}$, kWh
Jan	0.122446	0.528488	1.153071	0.194288
Feb	0.509767	1.039766	1.037586	0.95876
March	0.25141	0.892354	1.09223	1.05142
April	1.321673	1.172989	1.090559	0.852298
May	1.142417	1.328632	1.228063	1.362532
June	1.160591	1.315183	1.552475	1.12536
July	1.125002	1.669545	1.085193	1.034474
Aug	0.792334	1.314082	1.262762	0.857274
Sep	0.788575	0.983101	1.603325	1.474655
Oct	1.822925	0.553487	1.290697	1.216458
Nov	1.036426	0.942034	1.119736	0.517636
Dec	1.469121	0.624152	1.040096	1.049727

Table 3. Number of clear days fall in different weather condition for New Delhi weather station

Type of	Jan	Feb	Mar	April	May	June	July	Aug	Sep	Oct	Nov	Dec
A	3	3	5	4	4	3	2	2	7	5	6	3
B	8	4	6	7	9	4	3	3	3	10	10	7
C	11	12	12	14	12	14	10	7	10	13	12	13
D	9	9	8	5	6	9	17	19	10	3	2	8

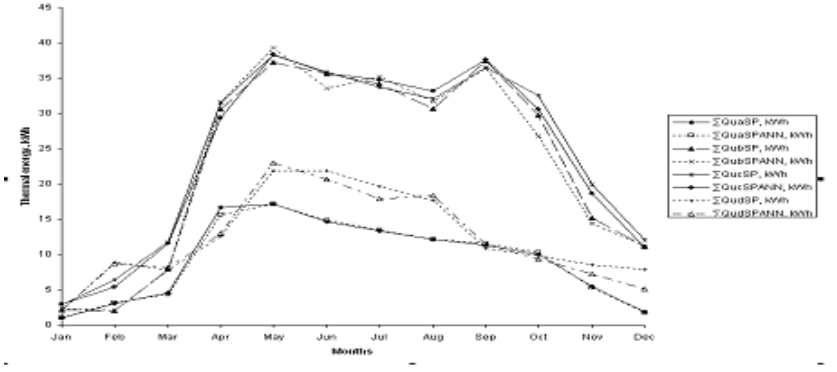


Fig. 5. Monthly variation of thermal energy for a, b, c and d type weather condition for New Delhi for HPVT-SPAC

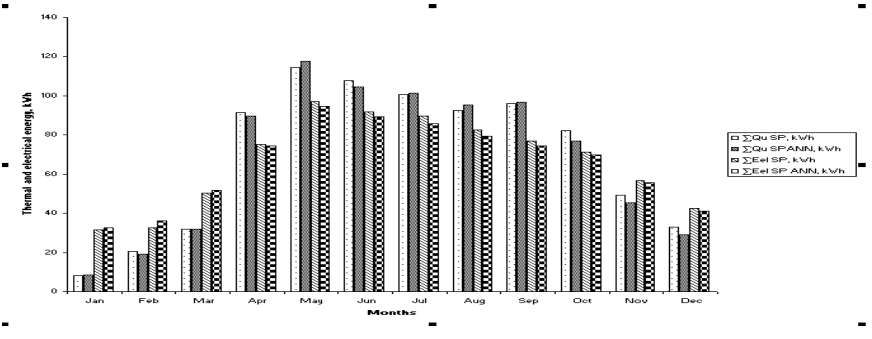


Fig. 6. Month variation of thermal and electrical energy for HPVT-SPAC

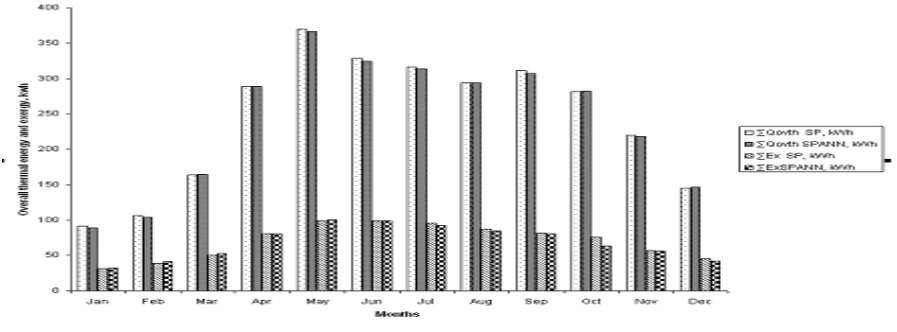


Fig. 7. Monthly variation of overall thermal energy and exergy for SPAC

6 Conclusion

In this paper an ANN model has been developed using MATLAB 7.1 neural networks toolbox for performance analysis of a semi transparent HPVT-SPAC. The ANN model is based on feed forward back propagation algorithm with 1 hidden layer. The LM with 12 neurons for semi transparent HPVT-SPAC in the hidden layer and 4 neurons in input and output layer is the most suitable algorithm with MSE value of 0.005. The RMSE varies from 0.122-1.822%, for different output parameters of semi transparent HPVT-SPAC respectively. There is a fair agreement between ANN and analytical values.

References

[1] Jiang, Y.: Prediction of monthly mean daily diffuse solar radiation using artificial neural networks and comparison with other empirical models. Energy Policy 36, 3833–3837 (2008)
[2] Leal, S.S., Tíba, C., Piacentini, R.: Daily UV radiation modeling with the usage of statistical correlations and artificial neural networks. Renew. Energy 36, 3337–3344 (2008)
[3] Koca, A., Oztop, H.F., Varol, Y., Koca, G.O.: Estimation of solar radiation using artificial neural networks with different input parameters for Mediterranean region of Anatolia in Turkey. Expert Systems with Applications 38(7), 8756–8762 (2011)
[4] Sencan, A., Ozdemir, G.: Comparison of thermal performances predicted and experimental of solar air collector. Journal Applied Science 7, 3721–3728 (2007)
[5] Caner, M., Gedik, E., Kecebas, A.: Investigation on thermal performance calculation of two type solar air collectors using artificial neural network. Expert Systems with Applications 38(3), 1668–1674 (2011)
[6] Singh, H.N., Tiwari, G.N.: Evaluation of cloudiness/haziness factor for composite climate. Energy 30, 1589–1601 (2005)
[7] Kamthania, D., Nayak, S., Tiwari, G.N.: Energy and Exergy Analysis of a Hybrid Photovoltaic Thermal Double Pass Air Collector. Applied Solar Energy 47(3), 199–206 (2011)
[8] Kamthania, D., Nayak, S., Tiwari, G.N.: Performance evaluation of a hybrid photovoltaic thermal double pass facade for space heating. Energy and Buildings 43(9), 2274–2281 (2011)

An Effective Software Implemented Data Error Detection Method in Real Time Systems

Atena Abdi[1], Seyyed Amir Asghari[1], Saadat Pourmozaffari[1], Hassan Taheri[2], and Hossein Pedram[1]

[1] Computer Engineering and Information Technology Department
[2] Electrical Engineering Department
Amirkabir University of Technology
Tehran, Iran
{atena_abdi,seyyed_asghari,saadat,pedram,saadat}@aut.ac.ir

Abstract. In this paper, a software-based technique is presented for detecting soft errors that damage data and values of the programs. The proposed technique that is called CPD (Critical Path Duplication) is based on critical path duplication of program. The mentioned path is extracted from the data flow graph of the program and has the most length so there is a great probability of error occurrence in it. In CPD technique, the instructions of the critical path is repeated and separated variables and registers are determined for them. In order to analyze the proposed technique, fault injection to variables and registers are utilized and the achieved results are compared with a full duplication method. Experimental results show that CPD technique has 54% and 25% of the performance and memory overheard less than full duplication method. The percentage of fault coverage is reduced about 24% which is acceptable in safety- critical applications which are sensitive to speed and space overheads.

Keywords: critical path duplication, error detection, fault injection, software redundancy.

1 Introduction

Soft errors that are important in computer systems influence on control flow or the data and variables of the program. A great percentage of transient faults in system are converted to ineffective and silent faults in the total output of the system; in the other words, their impact in system is deleted and not transferred to the final output. From the remained percentage, it is proved that about 33% to 77% are converted to Control Flow Errors (CFE) and remain percentage are converted to data errors [1].

Therefore, it can be concluded that replacing new techniques for detecting data and control flow errors (CFEs) instead of traditional technique of transient fault detection can eliminate the excessive cost of detecting ineffective errors of output from the system and so improve system performance and reduce its cost [2]. The first and the most important step in system fault tolerance against transient faults is these fault detection. Success in this stage can provide appropriate fault coverage for system.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 919–926.
springerlink.com

Several techniques have already been presented to control flow error detection and they are mainly based on program division into basic blocks and signature assigning to each block [2-5]. There are also many methods to data error detection that can be classified into two main types of hardware and software. The most common hardware technique for data errors detection is utilizing N parallel modules in order to do the operation and compare their results with each other. This technique has 100 (N-1) % memory and performance overhead and its fault coverage is about 100%. In many applications, hardware techniques are not acceptable because of their much overhead and so utilizing system level methods such as running a redundant task is more popular and prevalent for data error detection.

In order to decrease hardware techniques overheads, many different redundancies such as software redundancy, time redundancy, and information redundancy have already been utilized. Among them, software redundancy and replicating the program use is more popular. The reason of this diversity for the mentioned methods is that by the technology progress and reduction of electronic equipment dimensions, their vulnerability against transient faults and soft errors is increased. Therefore, delivering a method that is able to have good fault coverage by an acceptable overhead impose on memory and performance is very important. The parameters of memory consumption overhead, performance and speed overhead of the proposed technique, and fault coverage percentage are all important.

The proposed method of this paper considers these parameters altogether. Beside these limitations, the methods that create great redundancy in the memory are not appropriate for real time applications that meeting time demands are very important for them.

In order to data error detection, many methods have already been delivered which are based on software redundancy and duplication of all variables and program instructions. In these methods, variables are duplicated and the results of original and replica programs compared with each other and any mismatch reports an error. The mentioned techniques impose great memory and performance overhead to the system and for reducing this overhead, it is tried to eliminate comparison instruction to the extent that is possible and put them in necessary places. Some of the presented techniques in this field are delivered in [6-7].

In this paper, a pure software technique is proposed for data error detection that has much better performance and memory overhead in comparison with other implemented techniques. The proposed method of this paper, like other methods of this field, is based on software redundancy, but the duplication is not limited to the whole data and variables of the program and it repeats a path of data that is critical due to its importance.

In the second section of this paper, a review of the methods in this field is presented. CPD technique is broadly explained and analyzed in the third section. The forth section delivers experimental results and technique analysis. Finally, the results and a summary of the paper will be delivered.

2 Previous Works

Soft errors that are created due to heavy ion radiation, power supply distortion and other factors like this influence on flow and data of programs. Many techniques have been delivered to control flow error detection that are mainly software-based and

works by utilizing signature assigning and program division to basic blocks. Such methods are like [2] that their fault coverage is about 98%.

The remained percentage of soft errors leads to program data change. In order to detect such errors that consist 30% of all soft errors, many techniques have been delivered that are generally based on software redundancy and replication of running program variables and their main difference is in their comparison instruction place.

One of the most popular data error detection techniques is EDDI [6]. Instruction redundancy of this technique has no influence on the program output but detects errors while running. The main idea is instruction duplication in several registers and variables. The amounts of these two registers are compared with each other and one error is reported in the case of mismatch. Some comparison instructions are added at the end of computation instruction stream that check the output of main and redundant running which their places are very important. This comparison is exactly performed before store instructions or jumps in the memory.

Timing between redundant and main instructions in this group of methods is usually based on list scheduling algorithm [6].

ED^{4}I method [7] is a SIFT (Software Implemented Fault Tolerance) method that can detect transient errors by running two different programs but with data diversity that implement a functionality and their output comparison. Transient faults that cause on of the programs faulty can be detected by this method. Such transient errors are like those transient errors that occur in processors and bit-flips. Bit-flips is an unwanted change in memory cell states that is created due to different factors and SEUs (Single Event Upset) are one of the most important ones that their occurrence percentage in space environment is much more that other error models like MBU, SEL, SEGR, SEB, etc. For example, bit-flips that occur in program code section during running the program can change program behavior and lead to inaccurate results. By comparison of inaccurate results (caused by faulty program) and accurate results (caused by without fault program), fault existence (in this case, bit-flip) can be detected in one of the programs when both main and redundant programs produce inaccurate output. ED^{4}I technique can detect the fault by continuing program running till the output of the programs differs with each other.

The point that should be noted here is that this technique is not able to detect those errors which cause the program to be placed in an infinite loop and cannot exit it (usually this case happen when registers like IP face unwanted changes and the next instruction pointer of the program jumps to an inaccurate place).These errors are usually called Control Flow Error Program (the programs that interfere the accurate running of the program). For detecting these faults, techniques like watchdog timer can be utilized.

Another presented technique in data error detection extracts variable relations and divides variables into two groups: 1) Middle variables that are important for computation of other parameters and, 2) Final variables that do not participate in any other parameter computation.

After dividing the variables into two groups, all variables are duplicates and after each writing action in final variables, an instruction should be placed for the comparison of main and duplicated variables.

If there is any difference between the amounts of these two variables, a data error is detected. Figure1 shows the process of this method in a sample program. In this program, a, b, c variables are middle and d is the final one.

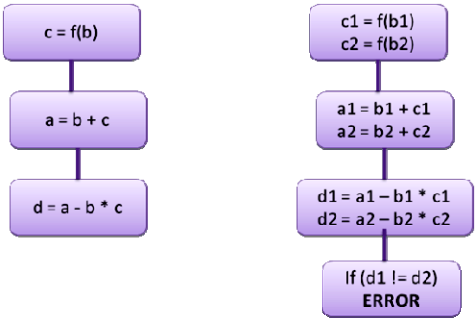


Fig. 1. a) usual running of the sample program b) sample program flow control after adding comparison and duplicated instructions

At the end of the computation, if final variables are not equal with each other, an error is reported. In this way, the comparison instruction is placed on a final variable only while writing and a great percentage of the occurred errors on program data are detectable and also redundant instruction overhead is reduced in comparison with similar methods [8].

The point that is similar in all mentioned methods is that they repeat the whole program that impose a great memory and performance overhead to the system, this problem is not acceptable in some applications like real time applications especially hard real time ones.

3 Description of the Proposed Technique

The purpose of presented method is to detect errors that effect on the data and variables of the program. In this method we use, data flow graph that can model the data and interconnection of every program [9]. In the data flow graph, nodes represent operands and the vertices represent the variables of the main structure of program. Figure2 shows a sample of this kind of graph.

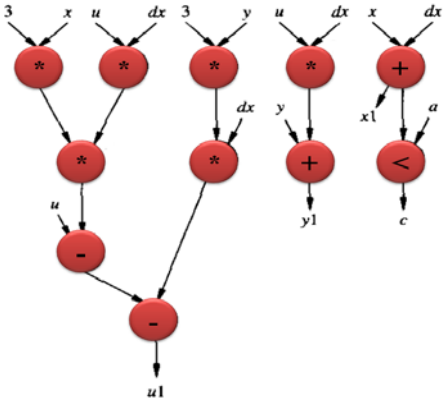


Fig. 2. Representation of a sample data flow graph

```
Diffeq{
read (x,y,u,dx,a);
        repeat{
                x1 = x + dx;
                u1 = u - (3*x*u*dx) - (3*y*dx);
                y1 = y + u * dx;
                c = x1 < a;
                x = x1; u = u1; y = y1;
                }
        until (c);
write (y);
}
```

Fig. 3. Differential equation solver code

The graph of Figure2 represents a sample differential equation solver code which is shown is Figure 3.

Using data flow graph, the interconnection of variables and their effects on each other is derived. Each data flow graph has a critical path which is the longest path of it. The longest path is very sensitive because a fault can propagate in it and damage lots of variables and operands. So the replication of this path can reduce the fault propagation and effect on the final result of program because it has some distributaries to other part of the graph and can affect them. If the critical path becomes fault tolerant, lots of data errors will be detected and the memory and performance overhead will be reduced.

In CPD (Critical Path Method), the data flow graph of each program and the critical path of it extracted according to ASAP and ALAP algorithm [9]. In the

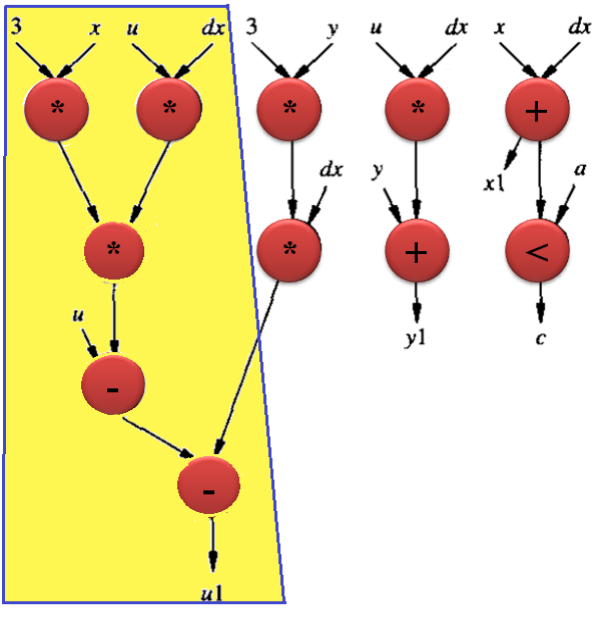


Fig. 4. The critical path of data flow graph of differential equation solver

mentioned technique, the graph schedules based on ASAP and ALAP algorithm, then if the results of them are identical for each node it is a part of critical path. The comprehensive description of critical path determination can be found in [9]. In CPD method, the instructions and variables of critical path become duplicated and comparison instruction is inserted after each write in final variables. In this way the overhead of critical path reduces about 50%.

Figure4 represents the critical path of data flow graph of Figure 2. The path that highlights in Figure4 is the longest and duplicating it can detect lots of errors with low memory and performance overhead.

Summarily the main idea of this method is duplication of critical path. Because it is the longest path of the graph and by duplicating the running in it, lots of errors can be tolerated. CPD method reduces the memory and performance overhead.

The experimental evaluation and comparing of this method with full duplication will be present in the next section.

4 Experimental Results

Figure5 shows the suggested method for evaluation of CPD technique. The test method consists of these elements:

- A background debug module that uses for programming and debugging. This module can be used for fault injection [10].
- phyCORE-MPC555 evaluation board
- A personal computer

BDM module injects faults directly to microprocessor registers. The benchmark of experiments is differential equation solver program that show in Figure3. We inject 784 faults to this program and its instructions, registers and variables. We also change the content of code and data segments randomly and evaluate the error detection capabilities of our method.

The fault injection method of previous section is used for evaluation of CPD. Table1 shows the results of fault injection. The duplication method of [8] has a lot of memory and performance overhead because it duplicates all of the instructions and variables of the program. Full duplication increases memory and performance loss. According to results of Table1, the CPD technique reduces the fault coverage about 24% but improve memory and performance overhead 54% and 25%.

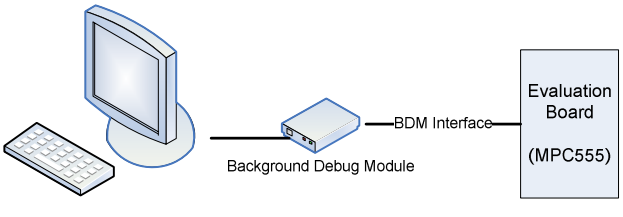


Fig. 5. Fault injection mechanism using BDM

Lots of applications use COTS equipments and their development board has specific MIPS. High percentage of memory overhead changes the MIPS and increases the amount power dissipation. So by using principle of locality we can add redundancy in critical regions and reduce overhead. In general purpose applications the sensitivity of operation is not as critical as its memory and performance overhead [11]. CPD method duplicates more important instructions and variables and reduces memory and performance overhead a lot.

Table 1. Fault coverage comparison of critical path duplication technique and full duplication method

Fault Coverage	Method
97.46%	Full Duplication
73.32%	CPD

Table2 shows memory and performance overhead of CPD and full duplication method. As this table shows CPD improve memory and performance overhead significantly.

Table 2. Comparison of mamory and performance overhead of CPD and full duplication methods

	Memory Overhead		Performance Overhead	
CPD	**Full Duplication**	**CPD**	**Full Duplication**	
1.03	1.38	0.6	1.3	**Diff equation solver**

5 Conclusion

Space and speed overhead, is so important in many real time applications because of their limited deadline and space. This paper presents a method based on software redundancy and replication of instructions for data error detection that calls CPD. In this method, only critical path instructions repeated and avoid fault propagation in this long path. Experiments show that CPD reduces performance and memory overhead 54% and 25% that is important in general purpose and real time applications very much in comparison to the most efficient full duplication method.

References

[1] Zhu, D., Aydin, H.: Reliability Effects of Process and Thread Redundancy on Chip Multiprocessors. In: Proc. of the 36th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (2006)

[2] Vemu, R., Abraham, J.: CEDA: Control-Flow Error Detection through Assertion. In: Proc. of the 12th IEEE international Online Testing Symposium (2006)

[3] Mahmood, A.: Concurrent Error Detection Using Watchdog Processors- A Survey. IEEE Transaction on Computers 37(2), 160–174 (1988)

[4] Oh, N., Shirvani, P.P., McCluskey, E.J.: Control-flow checking by software signatures. IEEE Transactions on Reliability 51(1), 111–122 (2002)
[5] Goloubeva, O., Rebaudengo, M., Reorda, M.S., Violante, M.: Softerror detection using control flow assertions. In: 18th IEEE International Symposium on Defect and Fault Tolerance in VLSI Systems (2003)
[6] Oh, N., Shirvani, P.P., McCluskey, E.J.: Error detection by duplicated instructions in super-scalar Processors. IEEE Trans. on Reliability 51(1), 63–75 (2002)
[7] Oh, N., Subhasish, M., McCluskey, E.J.: ED4I: Error detection by diverse data and duplicated instructions. IEEE Transaction on Computers 51(2), 180–199 (2002)
[8] Nicolescu, B., Velazco, R.: Detecting soft errors by a purely software approach: method, tools and experimental results. In: Proceedings of the Design, Automation and Test in Europe Conference and Exhibition (DATE 2003). IEEE Computer Society, Munich (2003)
[9] De Micheli, G.: Snthesis and optimization of digital circuits, 1st edn. McGraw-Hill Science, Engineering, Math. publication (1994)
[10] Asghari, S.A., Pedram, H., Taheri, H., Khademi, M.: A New Background Debug Mode Based Technique for Fault Injection in Embedded Systems. International Review on Modeling and Simulation (IREMOS) 3(3), 415–422 (2010)
[11] Shye, A., Blomstedt, J., Moseley, T., Janapa Reddi, V., Connors, D.: PLR: A Software Approach to Transient Fault Tolerance for Multi-Core Architectures. IEEE Transactions on Dependable and Secure Computing 6(2), 135–149 (2009)

Preprocessing of Automated Blood Cell Counter Data and Generation of Association Rules in Clinical Pathology

D. Minnie[1] and S. Srinivasan[2]

[1] Department of Computer Science, Madras Christian College, Chennai, India
minniearul@yahoo.com
[2] Department of Computer Science and Engineering, Anna University of Technology Madurai, Madurai, India
sriniss@yahoo.com

Abstract. This paper applies the preprocessing phases of the Knowledge Discovery in Databases to the automated blood cell counter data and generates association rules using apriori algorithm. The functions of an automated blood cell counter from a clinical pathology laboratory and the phases in Knowledge Discovery in Databases are explained briefly. Twelve thousand records are taken from a clinical laboratory for processing. The preprocessing steps of the KDD process are applied on the blood cell counter data. This paper applies the Apriori algorithm on the blood cell counter data and generates interesting association rules that are useful for medical diagnosis.

Keywords: Clinical Pathology, Blood Cell Counter, Knowledge Discovery in Databases, Data Mining, Association Rule Mining, Apriori algorithm.

1 Introduction and Related Work

The increased use of automated procedures in the health care industry has generated a huge volume of medical data which is generated from test results, patient details and details of treatment. This data can be effectively used to assist the health care professionals in efficient decision making.

Clinical Pathology is associated with monitoring diseases of patients by conducting tests on various body fluids. The fluids are either tested using manual procedure or an automated procedure. A Blood Cell Counter is an automated system that generates blood test results. The data contains noise such as missing values and the data is to be cleaned. The preprocessing phase of the Knowledge Discovery in Databases (KDD) Techniques is applied on the blood cell counter data to prepare the Blood Cell Counter Medical Data for efficient data mining. KDD [1], [2] is used to generate meaningful results from data and hence it is applied on medical data to generate knowledge.

Quality control is used in all laboratories to check errors and it plays a vital role in Clinical Pathology. The role of auto verification of results [6] in a laboratory information system is very important as the normal results can be generated at the speed of an the automated machine. The abnormal results have to be analyzed using various techniques. Specimen mislabeling is one of the errors present in Transfusion

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 927–935.
springerlink.com

Medicine and it can be reduced by collecting and trending the data on mislabeled samples with timely feedback to patient care [7].

Various combinations of Data Mining classification algorithms are used on medical data for efficient classification of data [8]. [9] presents ways of using sequences of clustering algorithms to mine temporal data. Association Rule Mining is used to diagnose diseases [10], [11] and risk patterns [12] from medical data. Taxonomy is used in certain cases to establish associations between different items in a data base [13]. Apriori algorithm is used to find frequent item sets in a database and to generate Association Rules from the frequent item sets [14]. A survey of various Data Mining Tools is presented in [15] and each of the tools is designed to handle a specific type of data and to perform a specific type of task.

Medical data is taken most of the times from medical records [16] and the data is found to be heterogeneous [17] in nature. The privacy issues [17] are to be finalized before handling medical data. The data that is taken from the Blood Cell Counter for our work is De-identified and the patient id and names are changed by the Clinical Pathology department before supplying the medical data for analysis.

2 Knowledge Discovery in Databases (KDD)

KDD consists of the processes Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Generation of Patterns and Knowledge Interpretation for effective knowledge generation and is shown in Fig.1.

In Data Cleaning the irrelevant data are removed from the collected data. In Data Integration multiple sources may be combined into a common source. The Data Selection process is involved with the selection of data relevant to the analysis and extracting them from the integrated data. The selected data is transformed to the appropriate form for the mining procedure. The process of extracting useful and implicit information from the transformed data is referred to as Data Mining. In Pattern Evaluation interesting patterns are identified from the processed data. The discovered knowledge is visually represented to the user in the Knowledge Representation process.

Data Mining is the Knowledge Discovery phase of KDD and it is the process of extracting implicit, useful, previously unknown, non-trivial information from data [1]. The techniques involved in Data Mining are grouped as Classification, Clustering, Association Rules and Sequences.

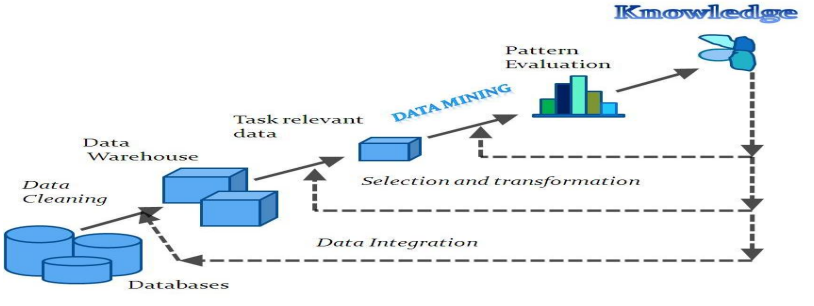


Fig. 1. Knowledge Discovery in Databases (KDD)

Classification is a supervised learning process and it maps data into known classes using Decision Trees, Neural Networks and Genetic Algorithms. Clustering is an unsupervised learning and it groups similar data into unknown clusters using K-Means, Nearest Neighbour and various other algorithms. Association Rule Mining (ARM) [4] uncovers relationships among data in a database.

Association Rule Mining (ARM) is used to find frequent patterns, associations and correlations among sets of items in databases and any other information repositories. Association Rule correlates the presence of one set of items with that of another set of items in the same transaction. The quality of an Association Rule is measured using its support and confidence values and several efficient methods are developed [5] to generate association rules.

Let X be an itemset with k elements $X_1, X_2, \ldots, X_k$ An Association Rule $X \rightarrow Y$ can be generated if the support of X and that of Y is above the minimum support value and also the confidence of the rule $X \rightarrow Y$ is above the minimum confidence specified.

Support S, of X is a probability that a transaction contains X and Support of Y is a probability that a transaction contains Y.

$$\text{Support } S(X) = P(X) \tag{1}$$

S(X) = Number of records with X/Total number of records

Also Support of $X \rightarrow Y$ is a probability that a transaction contains both X and Y.

$$\text{Support } S(X \rightarrow Y) = P(X \cup Y) \tag{2}$$

S (X→Y) = Number of records with both X and Y/Total number of records

Confidence C, of X→ Y is a conditional probability that a transaction that contains X contains Y also.

$$\text{Confidence } C(X \rightarrow Y) = P(X \mid Y) \tag{3}$$

C (X→Y) = Number of records with both X and Y /Number of records with X

$$C(X \rightarrow Y) = S(X \cup Y) / S(X) \tag{4}$$

ARM consists of the two major steps finding frequent itemsets and generating association rules. The itemsets which occur frequently in a database are called frequent itemsets. They have to satisfy the minimum support, min_sup specified. If the min_sup is 20%, then the itemsets whose support value is greater than or equal to 20 are considered as frequent itemsets. The set of all itemsets are called as candidate itemsets. Apriori algorithm uses the apriori principle "All subsets of an infrequent itemset are infrequent" and hence those subsets need not be considered for further processing. The association rules are generated using the frequent itemsets and they should satisfy the min_sup and a minimum confidence min_conf.

3 Automated Blood Cell Counter Data

The A Blood Cell Counter [3] is an automated machine that can be loaded with dozens of blood samples at a time and-the Complete Blood Count (CBC) or Full Blood Count (FBC) of the given blood samples are generated as a report. The number of red blood cells, white blood cells and platelets are some of the blood counts generated. The results are either printed directly or are stored in the computer for later use.

The 12,000 cell counter data are collected from a Clinical Pathology department of a reputed hospital. The data is present as an excel file and the data is used to generate association rules among the various attributes of the ABCC Database.

3.1 Automated Blood Cell Counter Data Format

The Blood Cell Counter Data is an excel file. A sample portion of it is given in fig. 2. The Blood Cell Counter generates files as output and the files consist of values for various attributes for each sample of blood.

The attributes of the records considered for further processing include PId, SId, PName, PAge, PGender, RDate, RTime, Hg count, MCH, MCHC, MCV, MPV, PCT and RDW and is shown along with a detailed description in Table 1.

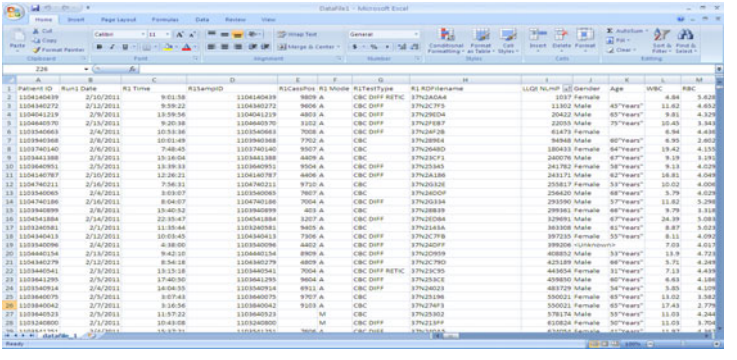


Fig. 2. Sample Blood Cell Counter Data

Table 1. Blood cell counter data attributes

Attribute Name	Attribute Description
PID	Patient Id
PNAME	Patient Name
PAGE	Patient Age
PGENDER	Patient Gender
SID	Sample Id
RD	Recorded Date
RT	Recorded Time
Hg	Hemoglobin Count
MCH	Mean Corpuscular Haemoglobin
MCHC	Mean Corpuscular Haemoglobin Concentration
MCV	Mean Corpuscular Volume
MPV	Mean Platelet Volume
PCT	Prothrombin Consumption Time
RDW	Red cell Distribution Width

These attributes are divided into two groups namely a group of attributes with constant range of normal values for all types of patients and different range for different types of patients. The MCV value has a single range of normal values whereas the Hg has different ranges of values for men, women and new born.

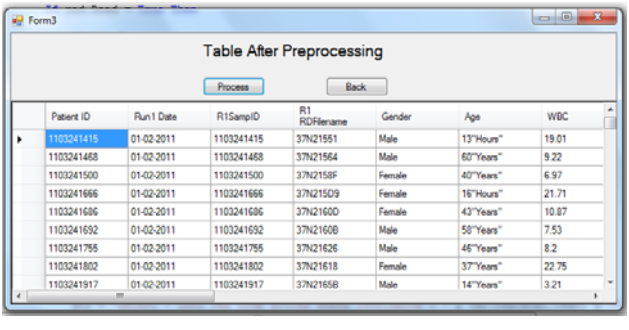
4 Results and Discussions

The Cell Counter Data was subjected to KDD processes from preprocessing to Data Mining to generate knowledge.

4.1 Data Cleaning

The process of detecting and correcting or removing corrupt or inaccurate records from a record set, table, or database is Data Cleaning. The Blood Cell Counter data contained missing fields and such records were not required for further analysis.

The attributes RDate, RTime, Hg count, MCH, MCHC, MCV, MPV, PCT and RDW were required for further processing and hence the records without these fields were removed. The resultant excel file contained the records with patient id, gender, age, date and time of results and the blood count fields were selected for further processing and is given in fig.3.



Patient ID	Run1 Date	R1SampID	R1 RDFilename	Gender	Age	WBC
1103241415	01-02-2011	1103241415	37N21551	Male	13"Hours"	19.01
1103241468	01-02-2011	1103241468	37N21564	Male	60"Years"	9.22
1103241500	01-02-2011	1103241500	37N2158F	Female	40"Years"	6.97
1103241666	01-02-2011	1103241666	37N215D9	Female	16"Hours"	21.71
1103241686	01-02-2011	1103241686	37N2160D	Female	43"Years"	10.87
1103241692	01-02-2011	1103241692	37N2160B	Male	58"Years"	7.53
1103241755	01-02-2011	1103241755	37N21626	Male	46"Years"	8.2
1103241802	01-02-2011	1103241802	37N21618	Female	37"Years"	22.75
1103241917	01-02-2011	1103241917	37N2165B	Male	14"Years"	3.21

Fig. 3. Preprocessed data

4.2 Data Selection

The cleaned excel data was taken as the data source to be used for data selection. The objective for the KDD process is to mine the preprocessed data to generate knowledge. The objective of the application of data mining process to the clinical data was to generate association rules that can be used for medical diagnosis.

The Clinical Pathologist requires the details such as Patient ID, Result Date, Result Time, Hgb, MDHC, MCH, MCV, MPV and PCT to check the quality of the system and hence they were considered for further processing.

4.3 Data Transformation

The Data Transformation stage is involved with converting data from a source data format into destination data format. The ranges of values for the attributes are used to find out whether the value is normal or abnormal. A value 1 is stored for the normal values and 0 is stored for the abnormal values. The flattened data is shown in fig 4.

4.4 Data Mining

The frequent item sets are identified from the medical data and the candidate 1 item sets are given in Fig.5, Frequent 1 –Itemsets are given in Fig. 6.

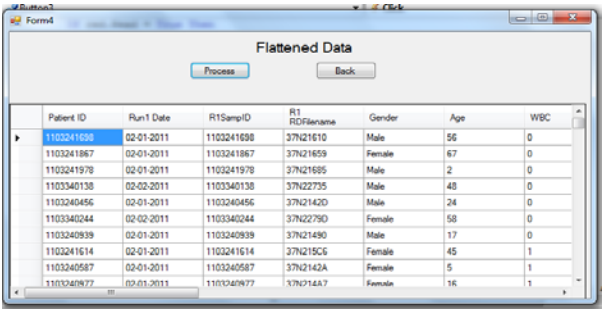


Fig. 4. Flattened data

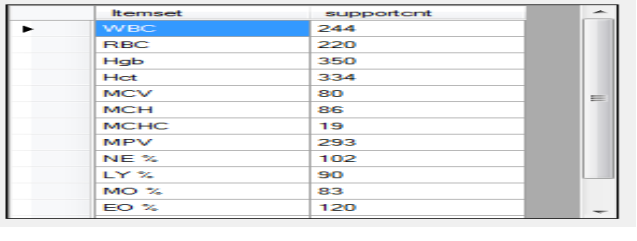


Fig. 5. Candidate 1- Itemsets

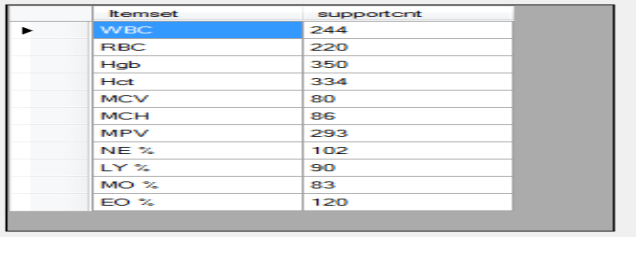
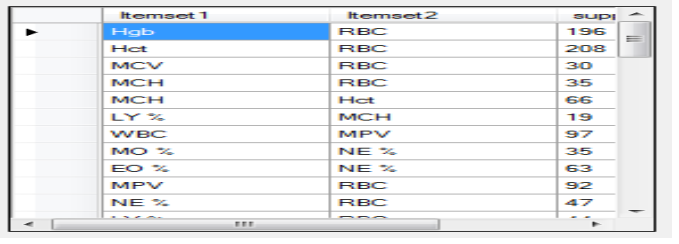


Fig. 6. Frequent 1 – Itemsets

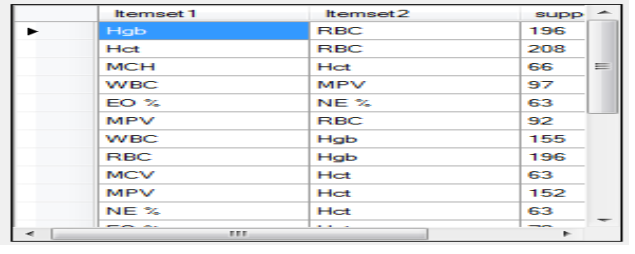
The frequent 1-item sets are joined to form candidate 2- Item sets and are shown in fig. 7 and the Frequent 2 – Item sets are shown in fig. 8.

The frequent 2-item sets are joined to form candidate 3- Item sets and are shown in fig. 9 and the Frequent 3 – Item sets satisfying the minimum support are shown in fig. 10.



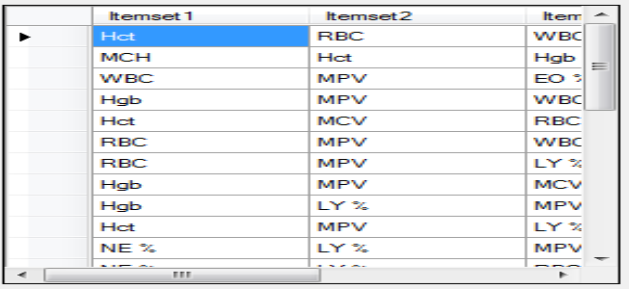
Itemset 1	Itemset 2	supi
Hgb	RBC	196
Hct	RBC	208
MCV	RBC	30
MCH	RBC	35
MCH	Hct	66
LY %	MCH	19
WBC	MPV	97
MO %	NE %	35
EO %	NE %	63
MPV	RBC	92
NE %	RBC	47

Fig. 7. Candidate 2 – Itemsets



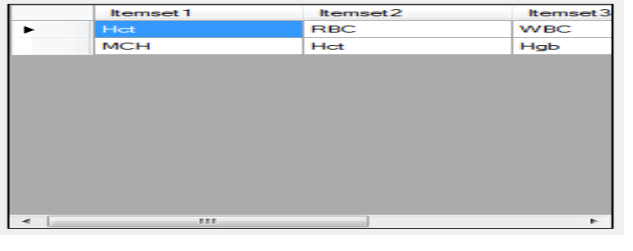
Itemset 1	Itemset 2	supp
Hgb	RBC	196
Hct	RBC	208
MCH	Hct	66
WBC	MPV	97
EO %	NE %	63
MPV	RBC	92
WBC	Hgb	155
RBC	Hgb	196
MCV	Hct	63
MPV	Hct	152
NE %	Hct	63

Fig. 8. Frequent 2 – Itemsets



Itemset 1	Itemset 2	Item
Hct	RBC	WBC
MCH	Hct	Hgb
WBC	MPV	EO %
Hgb	MPV	WBC
Hct	MCV	RBC
RBC	MPV	WBC
RBC	MPV	LY %
Hgb	MPV	MCV
Hgb	LY %	MPV
Hct	MPV	LY %
NE %	LY %	MPV

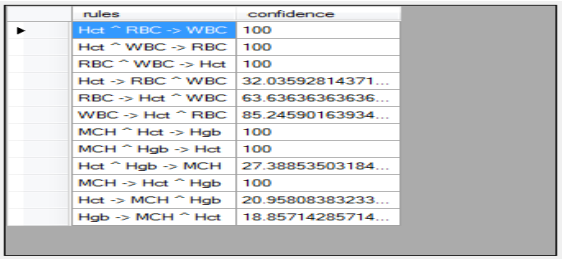
Fig. 9. Candidate 3 – Itemsets



Itemset 1	Itemset 2	Itemset 3
Hct	RBC	WBC
MCH	Hct	Hgb

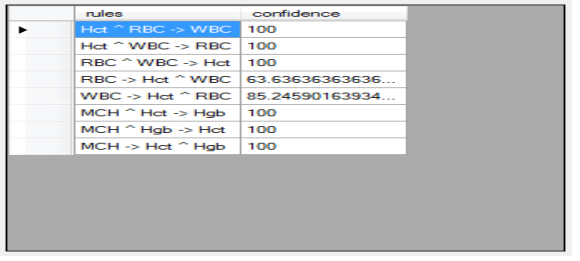
Fig. 10. Frequent 3 – Itemsets

The set of possible association rules are generated by applying the steps for identifying association rules from the frequent 3-item sets and are shown in Fig.11. The association rules satisfying the minimum confidence values are selected as the final association rules and are shown in fig.12.



rules	confidence
Hct ^ RBC -> WBC	100
Hct ^ WBC -> RBC	100
RBC ^ WBC -> Hct	100
Hct -> RBC ^ WBC	32.03592814371...
RBC -> Hct ^ WBC	63.63636363636...
WBC -> Hct ^ RBC	85.24590163934...
MCH ^ Hct -> Hgb	100
MCH ^ Hgb -> Hct	100
Hct ^ Hgb -> MCH	27.38853503184...
MCH -> Hct ^ Hgb	100
Hct -> MCH ^ Hgb	20.95808383233...
Hgb -> MCH ^ Hct	18.85714285714...

Fig. 11. Candidate Association Rules



rules	confidence
Hct ^ RBC -> WBC	100
Hct ^ WBC -> RBC	100
RBC ^ WBC -> Hct	100
RBC -> Hct ^ WBC	63.63636363636...
WBC -> Hct ^ RBC	85.24590163934...
MCH ^ Hct -> Hgb	100
MCH ^ Hgb -> Hct	100
MCH -> Hct ^ Hgb	100

Fig. 12. Generated Association Rules

5 Conclusion and Future Work

A brief study of Blood Cell Counter and Blood Cell Counter data is presented in the paper. The blood cell counter data was analyzed and few attributes were selected for processing, based on the knowledge given by the Clinical Pathologist. The KDD steps namely Data Cleaning, Integration, Selection, Transformation, and Mining were explained and were applied on the Blood Cell Counter Data to convert the raw data into a transformed data that was used for generating knowledge from the system. Frequent itemsets of the medical data are generated using apriori algorithm and association rules are generated.

Data from a single machine is used for this study and data from various Blood Cell Counter machines can be integrated for better knowledge generation.

Acknowledgements. The authors wish to thank Dr. Joy John Mammen, MD, Department of Transfusion Medicine and Immunohematology, Christian Medical College, Vellore, Tamilnadu, India for sharing his knowledge in Clinical Pathology, specially the functions of the Blood Cell Counter and also for providing the De-identified blood cell counter data.

References

[1] Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann Publishers (2006)

[2] Dunham, M.H.: Data Mining: Introductory and Advanced Topics. Pearson Education (2007)

[3] Automated Blood Cell Counter, http://www.medscape.com

[4] Goh, D.H., Ang, R.P.: An Introduction to Association rule mining: An application in counseling and help-seeking behavior of adolescents. Behaviour Research Methods 39(2), 259–266 (2007)

[5] Agrawal, R., Imielinski, T., Swami, A.: Mining Associations between Sets of Items in Large Databases. In: Proc. of the ACM-SIGMOD 1993 Int'l. Conference on Management of Data, pp. 207–216 (May 1993)

[6] Duca, D.J.: Auto Verification in a Laboratory Information System. Laboratory Medicine 33(1), 21–25 (2002)

[7] Quillen, K., Murphy, K.: Quality Improvement to Decrease Spe-cimen Mislabeling in Transfusion Medicine. Archives of Pathology and Laboratory Medicine 130, 1196–1198 (2006)

[8] Aslandogan Alp, Y., Mahajani, G.A.: Evidence Combination in Medical Data Mining. In: Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC 2004), vol. 2, pp. 465–469 (2004)

[9] Agrawal, R., Imielinski, T., Swami, A.: Database Mining: A Performance Perspective. IEEE Transactions on Knowledge and Data Engineering 5(6), 914–925 (1993)

[10] Toussi, M., Lamy, J.-B., Le Toumelin, P., Venot, A.: Using data mining techniques to explore physicians' therapeutic decisions when clinical guidelines do not provide recommendations: methods and example for type 2 diabetes. BMC Medical Informatics and Decision Making, 9–28 (2009)

[11] Dogan, S., Turkoglu, I.: Diagnosing Hyperlipidemia using Association rules. Mathematical and Computational Applications, Association for Scientific Research 13(3), 193–202 (2008)

[12] Li, J., Fu, A.W.-C., He, H., et al.: Mining risk Patterns in Medical data. In: KDD 2005, Chicago, Illinois, USA, pp. 770–775 (2005)

[13] Srikant, R., Agrawal, R.: Mining Generalized Association Rules. In: Proceedings of the 21st International Conference on Very Large Data Bases, Zurich, Swizerland (September 1995)

[14] Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules. In: Proceedings of the 20th International Conference on Very Large Data Bases, Santiago, Chile (September 1994)

[15] Goebel, M., Gruenwald, L.: A Survey of Data Mining and Knowledge Discovery Software Tools. In: SIGKDD Explorations, ACM SIGKDD (June 1999)

[16] Cerrito, P., Cerrito, J.C.: Data and Text Mining the Electronic Medical Record to Improve Care and to Lower Costs. In: Proceedings of SUGI 31, March 26-29, pp. 77–31 (2006)

[17] Cios, K.J., Moore, G.W.: Uniqueness of Medical Data Mining. Artificial Intelligence in Medicine 26(1-2), 1–24 (2002)

The Particular Approach for Personalised Knowledge Processing

Stefan Svetsky, Oliver Moravcik, Pavol Tanuska, Jana Stefankova, Peter Schreiber, and Pavol Vazan

Institute of Applied Informatics, Automation and Mathematics, The Faculty of Materials Science and Technology Slovak University of Technology, Trnava, Slovakia
stefan.svetsky@stuba.sk, oliver.moravcik@stuba.sk, pavol.tanuska@stuba.sk, jana.stefankova@stuba.sk, peter.schreiber@stuba.sk, pavol.vazan@stuba.sk

Abstract. Researchers, teachers, librarians, and individuals in daily practice perform various activities that require the processing of large amounts of knowledge and dynamic information flow. The database application BIKE (Batch Knowledge and Information Editor) was developed within the implementation of Technology Enhanced Learning. Based on the paradigm of "batch knowledge processing", it works as a universal, multi-purpose, pre-programmed "all-in-one" environment. The environment enables individuals to solve personalised approaches for producing teaching and learning materials, batch information retrieving, a personnel information system and also other applications. Some applications are presented in this paper.

Keywords: Technology enhanced learning, knowledge processing, engineering education, database applications, personalised learning environment.

1 Introduction

Information and knowledge play a key role in the research and educational activities. Despite the huge progress of Internet and Communication Technologies, the appropriate software and informatics tools on the market lack processing tailored to the individual user (researcher, teacher, and student). In the period 2000 – 2005, the database application Zápisník (WritingPad) was developed within an industrial research environment; it was built on the FoxPro for Windows database. This served for computer support of a research - development laboratory of surface treatments and some research projects (for example, the laboratory's information system was developed). After the modification of the WritingPad at the University, the processing of information was extended to the processing of knowledge. In this case, the power and advantages of conventional database technology of the nineties was used and was tailored for the processing of the huge amounts of data. During that time this advanced technology, could not be fully used by individuals due to the lower level of

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 937–946.
springerlink.com

ICT. However, the introduction of an entirely new paradigm of batch processing of information and knowledge was needed, because in a conventional DBMS (Database Management System) the data are processed in another way based on a relational model.

This gradually resulted in the development of the pre-programmed environment BIKE that was used for support of engineering education of bachelors. The existence of such an informatics tool allowed for teachers to solve the first stage of processing the knowledge flow between information sources and the "knowledge" database tables. User outputs led both to the BIKE environment and HTML - for-mat, which is readable by the common Internet browsers (browsable outputs). In this case, the default is set to Internet Explorer and Opera, and in some cases the browser Google Chrome is used. At this stage, a knowledge base and library with educational materials was created on an open web-domain and on the faculty server. This engineering content was processed using the BIKE or its selected standalone Zápisník / Writing Pad (file GeniusV.exe). The outputs supported a variety of educational activities (support for teacher's personal activities; for blended, informal, distance, active learning). After analyzing the state-of-the-art, the issue of the more or less empirical research was categorized into the field of Technology Enhanced Learning (originally the issue seemed to be eLearning) [1]. In the continuously published results the BIKE was presented as the informatics tool (an in-house software) which was designed in the Technology Enhanced Learning for processing of content [e.g. 2, 3, 4, 5].

2 Enlarged Approach for Technology Enhanced Learning

The processing of knowledge (engineering content) needs to address the flow of knowledge between the produced learning materials and the libraries (tailored for courses of study) and between individual educational activities. This represents the second phase of the solution for the flow of knowledge within which it was also automatically created a personalized virtual learning environment (VLE). Computer support of educational activities at this stage was more technology – driven than educational (pedagogical) – driven. Fig. 2.1 shows both stages of knowledge flow between information sources – knowledge tables (DBMS) and the teacher's activities.

The empirical research has showed that the work of university teachers in engineering knowledge is highly sophisticated. From an informatics perspective, this means that support of mental, learning activities and teaching is characterized by the use of unstructured information and knowledge. Moreover, the actual learning activities are unstructured and are not defined by their exact procedures. Teachers usually do this quite unconsciously. Thus, each type of learning activity should be studied, analysed, (repeatability, elements, action sequences), this is the basis of the program codes. This creates a way to gradually set a menu as optional and also as specific tools. Each menu item can be tested directly in teaching and if it is successful it can lead as an introduction to the default menu BIKE environment.

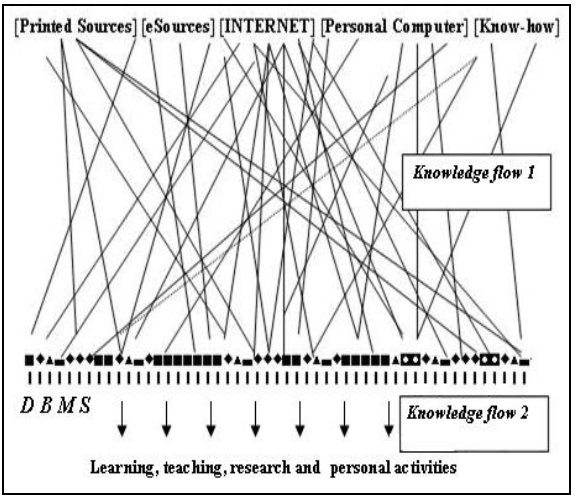


Fig. 2.1. Knowledge flow (information sources – BIKE – teacher's activities)

Two new elements come into play in practice. One is that the teacher selects different pedagogical approaches and teaching methods according to his needs to address communication with students (feedback). The second element is that it has to address the need to automate all kinds of normal teaching activities. This also applies to students, researchers and R&D staff in general. Therefore, the mechanical application of some existing general software (e.g. Learning Management System, Virtual Learning Environment, Educational Technology, Web 2.0) is not possible. It should also be taken into consideration, as already mentioned, for the need to promote the "automation" of mental activities. One can imagine this as an external chip, "Mind-ware", this is to say that technology is a partner of a person and creates a "social memory of individuals" linked to "global social memory" [see Saljö in 6]. As the teacher plays a key role technology must adapt. This is the principle of any automation solution. Computer aided education should be focused on those two elements on the level of educational quality and automation activities of the teacher and students. If we hold to Technology Enhanced Learning, the priority must be given to a pedagogically (Education) – driven. This principle expresses the scheme presented in Fig. 2.2.

From an informatics point of view it can be noted that the need for the above-mentioned processing of unstructured information, knowledge and activities, in this case, it is appropriate to use procedural programming (object-oriented and visual programming at this stage seems to be inappropriate). Conventional data-base environment FPW 2.6a running under all Windows operating systems makes it possible.

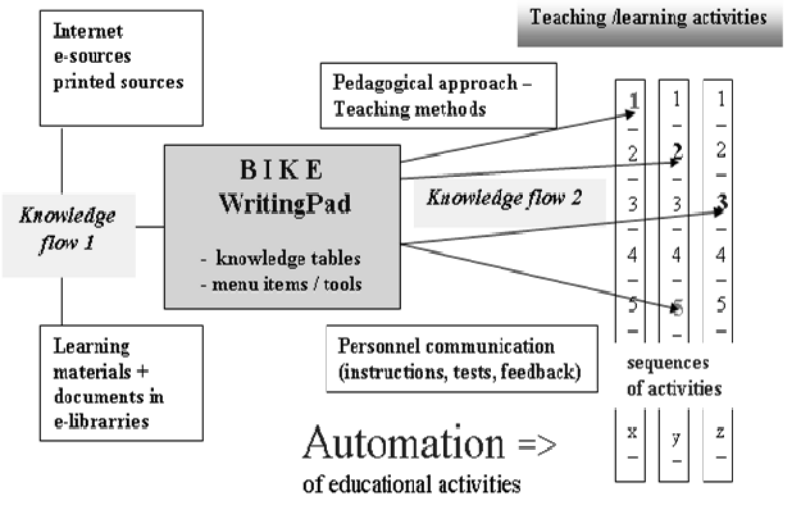


Fig. 2.2. Schema of the "Education-driven" Technology Enhanced Learning approach

Moreover, from the perspective of an individual it is very user friendly because it has its own text editor and allows one to create an exe – application. In addition, currently available databases that are commercial or free programs or office suites (MS Office, Open Office) are normally not suitable for the processing of unstructured knowledge. In the next stage of empirical research in engineering education, therefore, blocks of activities began to program which took into account both the pedagogical quality of teaching (pedagogical approach, didactic methods, communication) and pedagogically – driven automation. The programming codes gradually expand the menu of BIKE. It should be mentioned in particular that the solution to personal communications was in by adding the programming of the application PHP/MySQL. The application is launched and edited in the BIKE environment. It is installed on the computer of the teacher as well as on the faculty server. The combination of PHP with the MySQL database is used for activities in the on-line mode, i.e. mainly on personal social network programming between teacher and students.

A particularity of the integration of computer-aid into the engineering education is demonstrated by the fact that on one hand the teacher has the BIKE environment with optional information tools (menu items), and on the other hand, their use can not make students or does not enable the education itself or technical equipment of computer classes. For example, a teacher wants to improve the quality of teaching by introducing elements of support in technical English. However, the subject is only taught in their mother tongue. Thus, the application of programming itself is not sufficient. The difficult part is in how to figure out away to incorporate the teaching of (time, content, methodology). This experience was gained when dealing with the support for teaching technical English (a teacher has mastered the application WritingPad – Text To Speech Technology, but students did not use it). A need to harmonize informatics and pedagogical approaches is therefore needed.

Another specific example from the practice may be the implementation of batch internet retrieving. Here we have experienced such a paradox that a group of students to whom have been described and explained in great detail, a demonstration of how to work with the WritingPad on batch retrieving in collaboration with the Opera browser did not manage this activity. Conversely, it was managed by a group of students, even without the presence of a teacher. It was enough just to tell them: "Go to class on your computer, click the icon Genius in the top menu, click SvDopl2 and type the keywords into the box then see what is displayed on your interactive screen, be sure to first open Opera". Here was important also the fact that one or two students understood and explained to others.

3 Solution of Applications in Education on the Base of Automation

Practical solution of harmonisation of pedagogical – driven and technology driven approach in Technology Enhanced Learning (as illustrated in Fig. 2.2) is shown in detail in Fig. 3.1.

Below are descriptions that illustrate a solution of the informatics tools in common teaching activities undertaken and progress of activities which one may con-template on how they can be integrated into the education of bachelor students. The following pictures display some of the processing solutions of knowledge and automation representing pedagogical – informatics approach according to the scheme in Fig. 3.1.

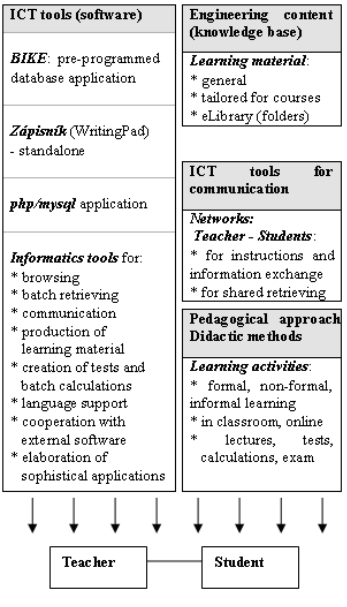


Fig. 3.1. Pedagogical – informatics background for solving the automation of teacher's / student activities

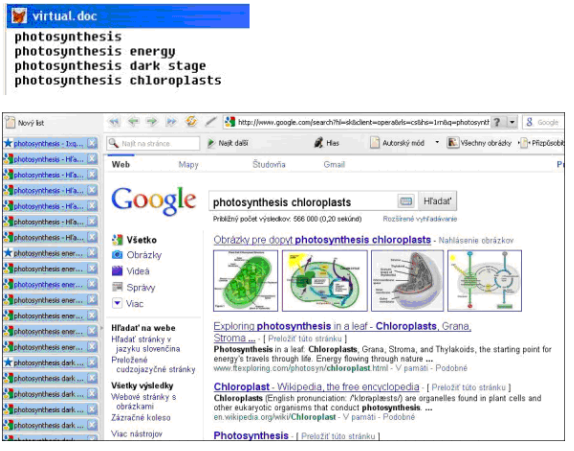


Fig. 3.2. Sample output from the batch internet retrieving of students

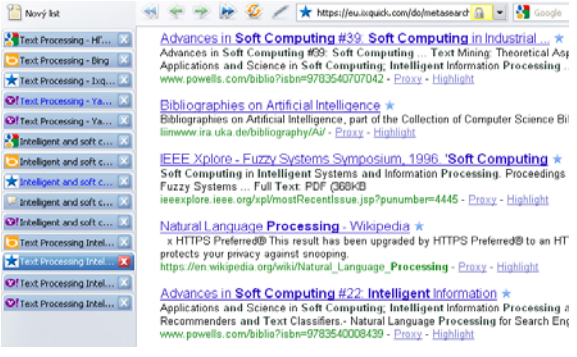


Fig. 3.3. Sample output from the batch internet retrieving of teacher

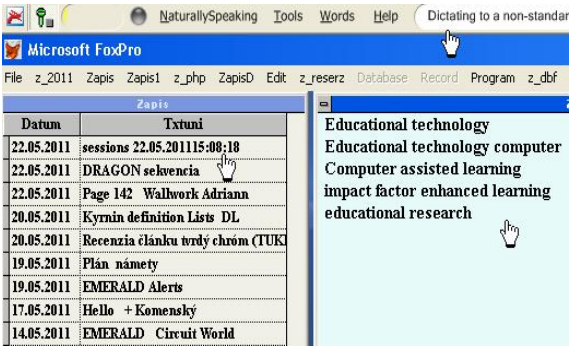


Fig. 3.4. Sample output of BIKE - space (a retrieving controlled by voice)

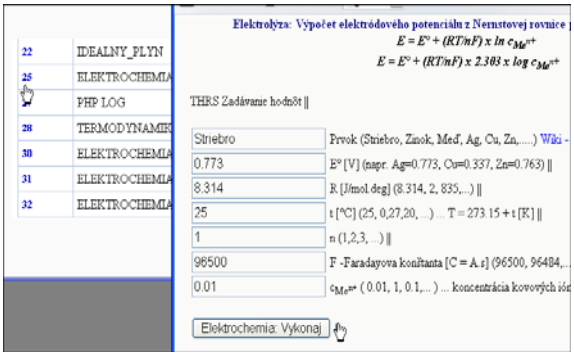


Fig. 3.5. Demonstration of tutorial of batch processed chemical calculations (Application of BIKE – PHP – MySQL from the server of the faculty)

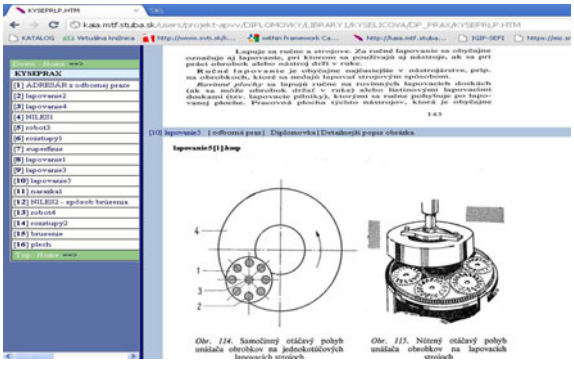


Fig. 3.6. Demonstration of solutions from diploma thesis – Secondary school study material for lapping (Z. Kyselicová)

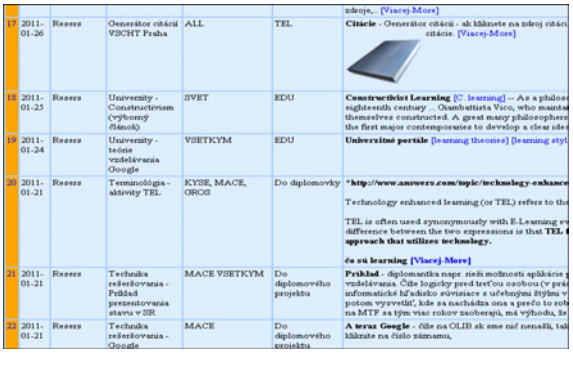


Fig. 3.7. Preview of personal social networks between teacher and students

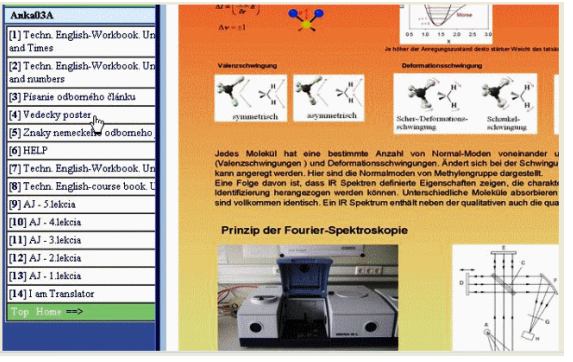


Fig. 3.8. Demonstration of the personal teacher support (Poster)

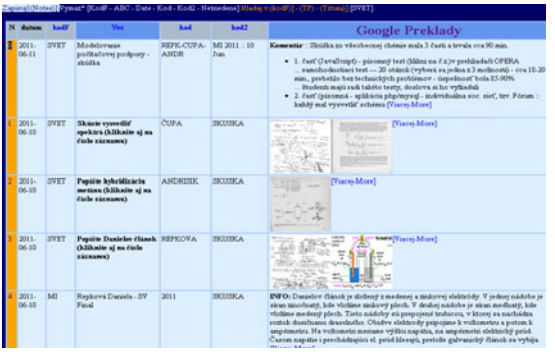


Fig. 3.9. Example of modeling a written examination on a personal social network between teacher and students of chemistry basics

4 Challenges for Knowledge Processing

The introduction of the pre-programmed BIKE environment to support engineering education created a basis for common personal support for activities of teachers, and students of bachelor studies. The teacher has a knowledge library (engineering content) and portfolio of informatics tools in the environment of BIKE. Based on the previous one it has a choice of how to support their daily activities (teaching, learning, publishing, other personal activities). Furthermore, when the real learning in the classroom takes place the previous can be combined where appropriate (e.g., learning material with self-evaluated tests and calculation exercises, etc.). The Standalone application WritingPad created from BIKE was introduced for use on computers in the classroom and was also used in some diploma theses. The theses showed that the future teachers of technical subjects can create e-learning materials using the WritingPad, to solve technological transfers from old books (based on digitization) or to support learning styles. Two students similarly designed small information systems within their technological focused theses. It has been shown that

the BIKE environment can produce user-friendly applications, i.e. the common informatics skills, software and hardware are sufficient. Moreover, ICT specialists can create sophisticated applications with this.

It should also be noted that if one does not start from the bottom-up concept of automation of the teacher's activities, the programmer using the technology – driven approach (top-down: technology to teachers) would often not even think that certain procedures must be programmed. This concept leads to the fact that in practice there can be completely new and unknown programming, because from an informatics point of view variants of support of mental activities should synergistic pedagogical – informatics outputs created. There does not exist a common system of be solved, and a set of informatics' tools, communication and educational activities proposed. Practically, this means that if we want to automatize even the simplest learning activities this will cause the creation of hundreds, thou-sands to an infinite number of solutions. Here it is also important for the harmonization with the Windows operating system, programming of which is often very onerous (BIKE is solved in such way so that it can use the maximum number of functions for which users are accustomed to, e.g., Explorer for off-line work).

5 Conclusions

The paper presented the universal approach for personalised knowledge processing. The key to the solution was the development of the pre-programmed BIKE multifunctional environment at the Faculty of Materials Science and Technology of e Slovak University of Technology. Based on the paradigm of batch knowledge processing, the environment works as a universal support tool, i.e. unlike other mono-purpose software solutions this enables individuals to concentrate information and knowledge at a given time and space within the multipurpose ""all-in-one" BIKE environment. Thus, individuals can create their teaching and learning materials, as well as a personnel information system, batch internet retrieval and also other personalised outcomes. Moreover, the environment is very user friendly and therefore it does not require any special informatics skills.

References

[1] European Commission TeLearn – European research on technology – enhanced learning. ICT Research in FP7 – Challenge 8: ICT for Learning and Access to Cultural Resources (2011), http://cordis.europa.eu/fp7/ict/telearn-digicult/telearn_en.html (accesed February 2011)

[2] Svetsky, S., et al.: Five years of research of technology enhanced learning implementation in teaching at the Faculty of Materials Science and Technology. Journal Research Papers MTF STU, 105–113 (2011) ISSN 1336-1589

[3] Svetsky, S., et al.: The Implementation of the Personalised Approach for Technology Enhanced Learning. In: WCECS 2010: Proceedings of the World Congress on Engineering and Computer Science 2010, San Francisco, USA, pp. 321–323 (2010) ISBN 978-988-17012-0-6

[4] Moravcik, O., et al.: Experiences with the Personalised Technology Support for Engineering Education. In: 21st Annual Conference of the Australasian Association for Engineering Education, Sydney, Australia, pp. 532–538 (2010) ISBN 978-0-646-54610-0
[5] Svetsky, S., Moravcik, O., Tanuska, P.: Some aspects of the technology enhanced learning in engineering education. In: Joint International IGIP-SEFI: Diversity Unifies - Diversity in Engineering Education, Trnava, Slovakia (2010) ISBN 978-2-87352-003-8
[6] Saljö, R.: Digital tools and challenges to institutional traditions of learning: technologies, social memory and the performative nature of learning (2010), doi:10.1111/j.1365-2729.2009.00341.x

Metrics Based Quality Assessment for Retrieval Ability of Web-Based Bioinformatics Tools

Jayanthi Manicassamy[1], P. Dhavachelvan[1], and R. Baskaran[2]

[1] Pondicherry University / Department of Computer Science, Pondicherry, India
jmanic2@yahoo.com
[2] Anna University / Department of CSE., Chennai, India

Abstract. Today, there is need for share and building resources that could be made available around the world any where, at any time. This made the necessity of web for which the usage and utilization depends on those who set these resources for sharing globally. Web based tool with the conceptual view of resource sharing it could be a classification of tool that only extracts resources from the web, like extraction of informatics from the database that would require which would be a thousand of thousands of object entities that would relay on this real world for which the resource is left for sharing globally. Bioinformatics tools aims at the same which is used for solving real world problems using DNA and amino acid sequences and related information using mathematical, statistical and computing methods. Mostly of the tools of this area are web-based since biological resources are real entity that should be kept updated based on the researches that requires vast space. Tools build in this area could not be build by one databases so, database like NCBI, PDB, EMBDL, Medline etc… have been developed to share its resources. At present development of bioinformatics tools are tremendously increasingly for real-time decision making for which it is vital to evaluate the performance of the tool by means of retrieval ability. Since mostly tools are web-based that utilizes various databases for information retrieval or mining information's it is vital for evaluating the retrieval ability along with error report of the tools for performance based quality assessment. Metrics is a measure that qualifies the characteristics of a product in numerical data that have being observed. In this paper few web-based bioinformatics tools have been taken, that retrieves documents from PubMed database for which the tools performances have been evaluated by quantitative means through metrics. Selective metrics that have been used are Information retrieval, error report, F-measure etc… for performance evaluation for which detailed result analysis have been narrated. From the observation made from the analyzed results on the tools will help to provide a guideline for developing better tools or selecting better tool for better decision making with enhanced qualities.

Index Terms: Bioinformatics, Databases, Information Retrieval, Performance Evaluation, Software Metrics, Web-Based.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 947–958.
springerlink.com

1 Introduction

Bioinformatics aim at making biological data computational in which various tools and methods have been developed and proposed in bioinformatics for automatic processing, which involves in discovering of new biological insights. Most of the tools developed utilizes web for various purpose that is based on the specific functionality for which the tool have been developed like biomedical literature [1] information extraction (IE) systems that incorporate natural language processing (NLP) techniques used in the biomedical field. Usage of ontology in bioinformatics as a means of accessing information automatically from large databases, for complex alignment of genomic sequences, in clustering of microarray DNA, pattern discovery [2] [5-11] etc… Biomedical uses various search techniques for mining literatures such as those offered by NCBI, PubMed system, require significant effort on the part of the searcher, and inexperienced searchers for using the systems effectively as experienced for easy and effective extraction [4] [6]. Today tremendous development of web-based tools and techniques are required to meet the demands in this area of bioinformatics.

Software metrics quantifies a product or process by means of measurement using numerical ratings. Statistical analysis is also being made to identifying the qualifying levels. Since most of the usages of the tools of this area are of web-based there is a need for evaluating the tools performance by means of its retrieval ability and error report. This is significant for exactly assessing the quality of the tool for better usage rather involving how well a system will perform in practical applications. This paper discusses and explores the main issues for evaluating retrieval ability, a key component in semantic web-based tools. A few selective web-based tools that retrieves data's from the same database have been taken for quantitative based qualitative assessment for which tool assessment is done practically by working on those tools. For each selected tool detailed analyses have been carried out and metrics have been identified that have been narrated for which evaluation has been carried out with detailed result analyses, which have been summarized along with graph representation in section 2.

2 Web-Based Tools Quantitative Evaluation

Several researchers are undergoing in the area of bioinformatics for which many approaches and tools are have been proposed mostly involved are of web-based. In this section we have explained about the selective web-based tools and Quantitative based quality assessment have made on those tools based on metrics for evaluating the retrieval ability performance.

A Web-Based Metrics

Evaluating tools by means of metrics is the major standard method adopted here for evaluating retrieval ability of the tools in this paper. We have identified a set of metrics $\{M_1, M_2 ... M_6\}$ for web-based tools evaluation which have been narrated below.

M1: Information Retrieval (IR). It measures the performance of retrieval ability of named entity recognition system which determines the system retrieval ability. It can be estimates as Information Retrieval (IR) = Number of data's correctly retrieved / Total number of data's retrieved [12].

M_2: F-measure. It measures the performance of named entity recognition system which determines the system stability. It can be estimates as F-measure = (2 * Precision * Recall) / (Precision + Recall) [13].

M_3: Lexical Overlap *(LO).* It measures the set of all domain relevant manual and extracted ontology. It can be estimates as LO = Correct / all [14].

M_4: Error Report. It is a global metrics used for measuring the percentage level of wrong output from the system for the given input. It can be estimates as Error Report = Percentage of wrong output from the system [1].

M_5: Lexical Precision (LPrecision). It is extraction precision metrics that measure the performance of the extraction modules through syntactic analyzes and pattern based extraction from the corpus. It can be estimates as *Lexical Precision (*LPrecision) = Correct $_{Extracted}$ / all $_{Extracted}$ [14]. Where, all $_{Extracted}$ is all the information that are being extracted and Correct $_{Extracted}$ is the correct information that are extracted from all the information that are being extracted.

M_6: Lexical Recall (LRecall). It measures the performance of the extraction modules through syntactic analyzes and pattern based extraction. It can be estimates as *Lexical Recall (*LRecall) = Correct $_{Extracted}$ / all $_{Corpus}$ [14]. Where, $Correct_{Extracted}$ are the correct information that are extracted from all the information that are being extracted and all_{Corpus} is all the information extracted that contains the defined keyword

B Quantitative Based Evaluation

Here, five web-based tools {wt_1, wt_2, wt_3, wt_4 and wt_5} have been taken for which four evaluations {e_1,e_2, e_3 and e_4} have been carried out on each tool. Four different inputs for evaluating the tools for which the evaluation input lies same for all tools have been given. The main reason behind is that it make easy tool evaluation and comparison for selection of the best decision making with the specific functionalities. For the evaluation carried out, results have been represented in Table 1. This evaluation is mainly for tools quality assessing which has been narrated in this section.

The evaluation scheme carried out for tools retrieval ability quality assessment mainly based on three metrics Information Retrieval, Recall and F-Measure for best analysis. IR represents performance of the system in the sense correct documents that are been accurately retrieved. High the IR the retrieval ability found to be high and the performance of the tool is high with comparison made with recall. The lower the recall the retrieval ability of relevant documents is considered to be less then even if retrieval ability is high the performance of the tool is considered to be less or better. F-Measure denotes the effectiveness of the tool and its variation based on the assessment carried out.

C Result Analysis

Result analysis has been made after each evaluation carried out on each tool and the over all evaluation of the tools has been represented as summary. The evaluated result value has been figured out in Table 1 for each evaluation the particular tool. Figure 1 represents the tools retrieval ability quality assessing in graph which is a visualized and easy mode of tools quality identification in the sense of retrieval ability based on analyzed results that have been summarized.

Jane Tool

Martijn J. Schuemie and Jan A. Kors [15] developed tool used to suggest journals and experts who have published similar articles to overcome difficulty in identifying appropriate journals to publish their work because many journals deal with multi-disciplinary.

Observation 1. "Breast Cancer" has been keyed as input for search to be made from PubMed journal database, 50 documents have been retrieved of which 45 found correct of which 47 found to contain the inputted keyword.

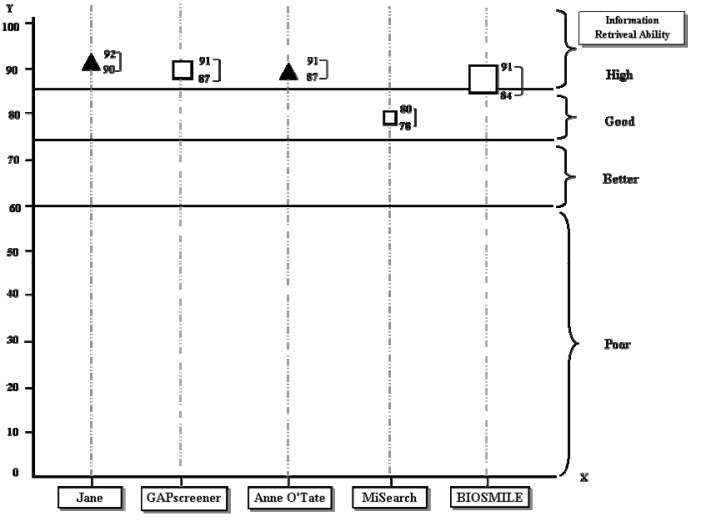
Proof. Consider e_1 for evaluating web-based tool (Wt_1) performance in Table 1. Based on the observation made metrics have been evaluated M_1 (IR) found to be 0.90. M_2 (F-measure) found to be 0.91, M_4 (Error Report) found to be 10 % or 0.10, M_5 (Precision) found to be 0.90 and M_7 (recall) found to be 0.95. Here the system found to have high retrieval ability based on the systems output. Thus the tool performance found to be high for this evaluation.

Observation 2. "Lung Tumor" has been keyed as input for search to be made from PubMed journal database, 54 documents have been retrieved of which 50 found correct of which 53 found to contain the inputted keyword.

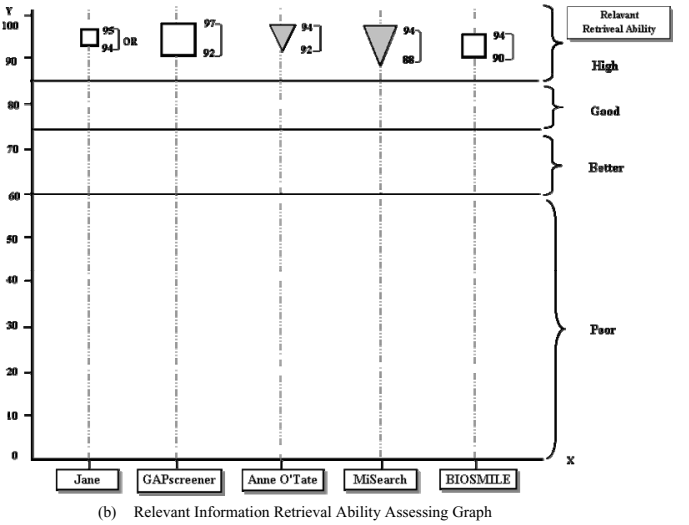
Proof. Consider e_2 for evaluating web-based tool (Wt_1) performance in Table 1. Based on the observation made metrics have been evaluated M_1 (IR) found to be 0.92. M_2 (F-measure) found to be 0.91, M_4 (Error Report) found to be 7.5 % or 0.075, M_5 (Precision found) to be 0.92 and M_7 (recall) found to be 0.94. Here the system found to have high retrieval ability and recall based on the systems output. Thus the tool performance found to be high for this evaluation.

Observation 3. "Flu Virus" has been keyed as input for search to be made from PubMed journal database, 51 documents have been retrieved of which 46 found correct of which 48 found to contain the inputted keyword.

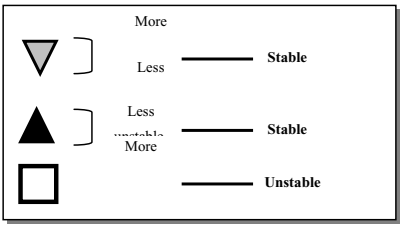
Proof. Consider e_3 for evaluating web-based tool (Wt_1) performance in Table 1. Based on the observation made metrics have been evaluated M_1 (IR) found to be 0.90. M_2 (F-measure) found to be 0.92, M_4 (Error Report) found to be 10 % or 0.10, M_5 (Precision found) to be 0.90 and M_7 (recall) found to be 0.95. Here the system found to have high retrieval ability and recall based on the systems output. Thus the tool performance found to be high for this evaluation.



(a) Information Retrieval Ability Assessing Graph



(b) Relevant Information Retrieval Ability Assessing Graph



where,

Stable – Represents the value will fall as the level of the surface.

Unstable – Represents the value range could not be predicted could vary within the stipulate range

Fig. 1. Tools Retrieval Ability Quality Assessing Graph

Overall Result Analysis: Based on the three evaluation carried out on the tool Information Retrieval falls from 0.90 to 0.92 where as recall is 0.94 or 0.95. Since Information Retrieval of the tool is high with high relevant documents retrieves the performance of the tool is high. Since F-measure is less in variation 0.02 the effectiveness of the tools is also high with high retrieval ability. Thus tool is of high quality.

GAPscreener Tool

Wei Yu et. al. [16] SVM-based tool made available for screening the human genetic association literature in PubMed has been used significantly to train the SVM model.

Observation 1. "Breast Cancer" has been keyed as input for search to be made from 01-01-2009 till 05-01-2009 from the PubMed for which 4557 document have been retrieved of EDAT data type of which 3985 found to be correct of which 4315 found to contain the inputted keyword.

Proof. Consider e_1 for evaluating web-based tool (Wt_2) performance in Table 1. Based on the observation made metrics have been evaluated M_1 (IR) found to be 0.87. M_2 (F-measure) found to be 0.89, M_4 (Error Report) found to be 12.6 % or 0.126, M_5 (Precision found) to be 0.87 and M_7 (recall) found to be 0.92. Here the system found to have high retrieval ability and recall based on the systems output. Thus the tool performance found to be high for this evaluation.

Observation 2. "Lung Tumor" has been keyed as input for search to be made from 01-01-2009 till 05-01-2009 from the PubMed for which 2205 document have been retrieved of EDAT data type of which 1978 found to be correct of which 2098 found to contain the inputted keyword.

Proof. Consider e_2 for evaluating web-based tool (Wt_2) performance in Table 1. Based on the observation made metrics have been evaluated M_1 (IR) found to be 0.89. M_2 (F-measure) found to be 0.91, M_4 (Error Report) to be 9.9 % or 0.099, M_5 (Precision found) to be 0.89 and M_7 (recall) found to be 0.94. Here the system found to have high retrieval ability and recall based on the systems output. Thus the tool performance found to be high for this evaluation.

Observation 3. "Flu Virus" has been keyed as input for search to be made from 01-01-2009 till 05-01-2009 from the PubMed for which 2205 document have been retrieved of EDAT data type of which 1989 found to be correct of which 2118 found to contain the inputted keyword.

Proof. Consider e_3 for evaluating web-based tool (Wt_2) performance in Table 1. Based on the observation made metrics have been evaluated M_1 (IR) found to be 0.90. M_2 (F-measure) found to be 0.91, M_4 (Error Report) to be 10 % or 0.10, M_5 (Precision found) to be 0.90 and M_7 (recall) found to be 0.93. Here the system found to have high retrieval ability and recall based on the systems output. Thus the tool performance found to be high for this evaluation.

Overall Result Analysis: Based on the three evaluation carried out on the tool Information Retrieval falls from 0.87 to 0.91 where as recall falls from 0.92 to 0.97. Since Information Retrieval of the tool is high with high relevant documents retrieves the performance of the tool is high. Since F-measure is moderate in variation 0.05 the effectiveness of the tools is also good with high retrieval ability. Thus tool is of high quality.

Anne O'Tate Tool

Neil R Smalheiser et. al. [17] developed for article processing that shows multiple aspects on articles retrieved from PubMed based on user pre-defined categories. It is a generic tool for summarization, of user friendly which drill-down and browsing of PubMed search results that accommodates a wide range of biomedical users and needs.

Observation 1. "Breast Cancer" has been keyed as input for search to be made from PubMed database for which 200793 document have been retrieved of which 180579 found to be correct of which 195582 found to contain the inputted keyword.

Proof. Consider e_1 for evaluating web-based tool (Wt_3) performance in Table 1. Based on the observation made metrics have been evaluated M_1 (IR) found to be 0.89. M_2 (F-measure) found to be 0.90, M_4 (Error Report) to 11 % or 0.11 and M_7 (recall) found to be 0.92. Here the system found to have high retrieval ability and recall based on the systems output. Thus the tool performance found to be high for this evaluation.

Observation 2. "Lung Tumor" has been keyed as input for search to be made from PubMed database for which 200793 document have been retrieved of which 140587 found to be correct of which 149827 found to contain the inputted keyword.

Proof. Consider e_2 for evaluating web-based tool (Wt_3) performance in Table 1. Based on the observation made metrics have been evaluated M_1 (IR) found to be 0.89. M_2 (F-measure) found to be 0.91, M_4 (Error Report) to 10.7 % or 0.107 and M_7 (recall) found to be 0.94. Here the system found to have high retrieval ability and recall based on the systems output. Thus the tool performance found to be high for this evaluation.

Observation 3. "Flu Virus" has been keyed as input for search to be made from PubMed database for which 15362 document have been retrieved of which 13987 found to be correct of which 14728 found to contain the inputted keyword.

Proof. Consider e_3 for evaluating web-based tool (Wt_3) performance in Table 1. Based on the observation made metrics have been evaluated M_1 (IR) found to be 0.91. M_2 (F-measure) found to be 0.92, M_4 (Error Report) to 9 % or 0.09 and M_7 (recall) found to be 0.94. Here the system found to have high retrieval ability and recall based on the systems output. Thus the tool performance found to be high for this evaluation.

Overall Result Analysis: Based on the three evaluation carried out on the tool Information Retrieval falls from 0.87 to 0.91 where as recall falls from 0.92 to 0.94. Since Information Retrieval of the tool is high with high relevant documents retrieves

the performance of the tool is high. Since F-measure is less in variation 0.03 the effectiveness of the tools is also high with high retrieval ability. Thus tool is of high quality.

MiSearch Tool

David J. States et. al. biomedical literature search tool [18] displays literatures based on rank citations that work on likelihood of the user involving statistical model. Based on dynamically updated users likelihood citation selections are automatically acquired during browsing.

Observation 1. "Breast Cancer" has been given as input out of 170454 citations retrieved from PubMed only top 5000 citations have been displayed of which 3875 found to be correct of which 4195 found to contain the inputted keyword.

Proof. Consider e_1 for evaluating web-based tool (Wt_4) performance in Table 1. Based on the observation made metrics have been evaluated M_1 (IR) found to be 0.78. M_2 (F-measure) found to be 0.84, M_5 (Precision found) to be 0.79, M_4 (Error Report) to 22.5 % or 0.225 and M_7 (recall) found to be 0.92. Here the system found to have good retrieval ability and high recall based on the systems output. Thus performance of the tool found to be high for this evaluation since the retrieval ability of relevant document is high rather than considering the retrieval ability of related documents which is found to be good.

Observation 2. "Lung Tumor" has been given as input out of 138598 citations retrieved from PubMed only top 5000 citations have been displayed of which 3995 found to be correct of which 4212 found to contain the inputted keyword.

Proof. Consider e_2 for evaluating web-based tool (Wt_4) performance in Table 1. Based on the observation made metrics have been evaluated M_1 (IR) found to be 0.79. M_2 (F-measure) found to be 0.86, M_5 (Precision found) to be 0.79, M_4 (Error Report) to 20.6 % or 0.206 and M_7 (recall) found to be 0.94. Here the system found to have good retrieval ability and high recall based on the systems output. Thus performance of the tool found to be high for this evaluation since the retrieval ability of relevant document is high rather than considering the retrieval ability of related documents which is found to be good.

Observation 3. "Flu Virus" has been given as input out of 138598 citations retrieved from PubMed only top 5000 citations have been displayed of which 4018 found to be correct of which 4515 found to contain the inputted keyword.

Proof. Consider e_3 for evaluating web-based tool (Wt_4) performance in Table 1. Based on the observation made metrics have been evaluated M_1 (IR) found to be 0.80. M_2 (F-measure) found to be 0.83, M_5 (Precision found) to be 0.80, M_4 (Error Report) to 19.7% or 0.197 and M_7 (recall) found to be 0.88. Here the system found to have good retrieval ability and high recall based on the systems output. Thus performance of the tool found to be high for this evaluation since the retrieval ability of relevant

document is high rather than considering the retrieval ability of related documents which is found to be good.

Overall Result Analysis: Based on the three evaluation carried out on the tool Information Retrieval falls from 0.78 to 0.80 where as recall falls from 0.88 to 0.94. Since Information Retrieval of the tool is good with high relevant documents retrieves the performance of the tool is good. Since F-measure is less in variation 0.03 the effectiveness of the tools is also high with good retrieval ability. Thus tool is of good quality.

Table 1. Web-Based Tools Metrics Evaluation

Metrics		Bioinformatics Web-Based Tools														
		Jane			GAPscreener			Anne O'Tate			MiSearch			BIOSMILE		
		Wt_1			Wt_2			Wt_3			Wt_4			Wt_5		
		e_1	e_2	e_3	e_1	e_2	e_3	e_1	e_2	e_3	e_1	e_2	e_3	e_1	e_2	e_3
M_1	Information Retrieval (IR)	0.90	0.92	0.90	0.87	0.89	0.90	0.89	0.89	0.91	0.78	0.79	0.80	0.84	0.89	0.91
M_2	F-Measure	0.91	0.93	0.92	0.89	0.91	0.91	0.90	0.91	0.92	0.84	0.86	0.83	0.86	0.91	0.90
M_3	Lexical Overlap							0.89	0.89	0.91						
M_4	Error Report	0.10	0.075	0.10	0.126	0.099	0.10	0.11	0.107	0.9	0.225	0.206	0.197	0.155	0.107	0.09
M_5	Lexical Precision	0.90	0.92	0.90	0.87	0.89	0.90				0.79	0.79	0.80	0.84	0.89	0.91
M_6	Lexical Recall	0.95	0.94	0.95	0.92	0.94	0.93				0.92	0.94	0.88	0.91	0.93	0.93

where,

Shaded cells represent the related metrics for that cell is not applicable for that tool.

BIOSMILE Tool

Hong-Jie Dai et. al. biomedical articles analyzation tool [19] based on selected verbs and user-defined relational information's. Users can select articles for further analysis that has been viewed as the abstract text or in table form. Here unique features are semantic relation analysis of abstracts and PPI relevance ranking for abstracts.

Observation 1. "Breast Cancer" has been keyed as input for search to be made from PubMed database for which 200793 document have been retrieved of which 169692 found to be correct of which 185979 found to contain the inputted keyword.

Proof. Consider e_1 for evaluating web-based tool (Wt_5) performance in Table 1. Based on the observation made metrics have been evaluated M_1 (IR) found to be 0.84. M_2 (F-measure) found to be 0.86, M_4 (Error Report) to 15.5 % or 0.155 and M_7 (recall)

found to be 0.91. Here the system found to have good retrieval ability and high recall based on the systems output. Thus performance of the tool found to be high for this evaluation since the retrieval ability of relevant document is high rather than considering the retrieval ability of related documents which is found to be good.

Observation 2. "Lung Tumor" has been keyed as input for search to be made from PubMed database for which 157369 document have been retrieved of which 140587 found to be correct of which 149979 found to contain the inputted keyword.

Proof. Consider e_2 for evaluating web-based tool (Wt_5) performance in Table 1. Based on the observation made metrics have been evaluated M_1 (IR) found to be 0.89. M_2 (F-measure) found to be 0.91, M_4 (Error Report) to 10.7 % or 0.107 and M_7 (recall) found to be 0.93. Here the system found to have high retrieval ability and recall based on the systems output. Thus the tool performance found to be high for this evaluation.

Observation 3. "Flu Virus" has been keyed as input for search to be made from PubMed database for which 15362 document have been retrieved of which 13987 found to be correct of which 14978 found to contain the inputted keyword.

Proof. Consider e_3 for evaluating web-based tool (Wt_5) performance in Table 1. Based on the observation made metrics have been evaluated M_1 (IR) found to be 0.91. M_2 (F-measure) found to be 0.90, M_4 (Error Report) to 9 % or 0.09 and M_7 (recall) found to be 0.93. Here the system found to have high retrieval ability and recall based on the systems output. Thus the tool performance found to be high for this evaluation.

Overall Result Analysis: Based on the three evaluation carried out on the tool Information Retrieval falls from 0.84 to 0.91 where as recall falls from 0.91 to 0.94. Since Information Retrieval of the tool is high with high relevant documents retrieves the performance of the tool is good. Since F-measure is moderate in variation 0.05 the effectiveness of the tools is also good with high retrieval ability. Thus tool is of high quality.

3 Conclusion

The main aim of the paper is to provide a set of guidelines for both retrieval ability and performance assessing of web-based tools in bioinformatics. This experiment has been designed such that the evaluation has done based on the evaluating metrics that are applicable for each tool in the path of assessing only the retrieval ability and performance for quality assessing. The metrics set used here for the tool evaluation provides a standard procedure for evaluating web-based tools. The results observed provided an idea in depth for evaluating retrieval ability of web-based bioinformatics tools. This also provided a way to identify the requirements that are required in the development of new tools in bioinformatics to overcome the pitfalls of the existing tools and provides a way for usage of the existing tools the provides better retrieval ability and performance.

References

[1] Witte, R., Baker, C.J.: Towards A Systematic Evaluation of Protein Mutation Extraction Systems. Journal of Bioinformatics and Computational Biology (JBCB) 1(6), 1339–1359 (2007)

[2] Vengattaraman, T., Abiramy, S., Dhavachelvan, P., Baskaran, R.: An Application Perspective Evaluation of Multi-Agent System in Versatile Environments. International Journal on Expert Systems with Applications 38(3), 1405–1416 (2011)

[3] Manicassamy, J., Dhavachelvan, P.: Automating diseases diagnosis in human: A Time Series Analysis. In: Proceedings of International Conference and Workshop on Emerging Trends in Technology (ICWET 2010), pp. 798–800 (2010)

[4] Manicassamy, J., Dhavachelvan, P.: Metrics Based Performance control Over Text Mining Tools in Bio-Informatics. In: ACM International Conference on Advances in Computing, Communication and Control, ICAC3 2009, pp. 171–176 (2009)

[5] Venkatesan, S., Dhavachelvan, P., Chellapan, C.: Performance analysis of mobile agent failure recovery in e-service applications. International Journal of Computer Standards and Interfaces 32(1-2), 38–43 (2005)

[6] Dhavachelvan, P., Uma, G.V., Venkatachalapathy, V.S.K.: A New Approach in Development of Distributed Framework for Automated Software Testing Using Agents. International Journal on Knowledge –Based Systems 19(4), 235–247 (2006)

[7] Dhavachelvan, P., Uma, G.V.: Complexity Measures For Software Systems: Towards Multi-Agent Based Software Testing. In: Proceedings - 2005 International Conference on Intelligent Sensing and Information Processing, ICISIP 2005, pp. 359–364, Art. no. 1529476 (2005)

[8] Dhavachelvan, P., Uma, G.V.: Reliability Enhancement in Software Testing – An Agent-Based Approach for Complex Systems. In: Das, G., Gulati, V.P. (eds.) CIT 2004. LNCS, vol. 3356, pp. 282–291. Springer, Heidelberg (2004)

[9] Dhavachelvan, P., Uma, G.V.: Multi-agent Based Integrated Framework for Intra-class Testing of Object-Oriented Software. In: Yazıcı, A., Şener, C. (eds.) ISCIS 2003. LNCS, vol. 2869, pp. 992–999. Springer, Heidelberg (2003)

[10] Victer Paul, P., Saravanan, N., Jayakumar, S.K.V., Dhavachelvan, P., Baskaran, R.: QoS enhancements for global replication management in peer to peer networks. Future Generation Computer Systems 28(3), 573–582 (2012)

[11] Victer Paul, P., Vengattaraman, T., Dhavachelvan, P.: Improving efficiency of Peer Network Applications by formulating Distributed Spanning Tree. In: Proceedings - 3rd International Conference on Emerging Trends in Engineering and Technology, ICETET 2010, pp. 813–818, Art. no. 5698439 (2010)

[12] Oberto, J.: BAGET: a web server for the effortless retrieval of prokaryotic gene context and sequence. ACM Portal, 424–425 (2007)

[13] Falchi, M., Fuchsberger, C.: Jenti: an efficient tool for mining complex inbred genealogies. Bioinformatics Oxford Journal 24(5), 724–726 (2008)

[14] Sabou, M., Wroe, C., Goble, C., Mishne, G.: Learning domain ontologies for web service descriptions: An experiment in bioinformatics, citeseer (2005)

[15] Schuemie, M.J., Kors, J.A.: Jane: suggesting journals, finding experts. Oxford Journal, 727–728 (2008)

[16] Yu, W., Clyne, M., Dolan, S.M., Yesupriya, A., Wulf, A., Liu, T., Khoury, M.J., Gwinn, M.: GAPscreener: An automatic tool for screening human genetic association literature in PubMed using the support vector machine technique. BMC Bioinformatics 9(1), 205 (2008)

[17] Smalheiser, N.R., Zhou, W., Torvik, V.I.: Anne O'Tate: A tool to support user-driven summarization, drill-down and browsing of PubMed search results. Journal of Biomedical Discovery and Collaboration 3(2) (2008)
[18] States, D.J., Ade, A.S., Wright, Z.C., Bookvich, A.V., Athey, B.D.: MiSearch adaptive pubMed search tool. Bioinformatics Oxford Journal, 974–976 (2008)
[19] Dai, H.-J., Huang, C.-H., Lin, R.T.K., Tsai, R.T.-H., Hsu, W.-L.: BIOSMILE web search: a web application for annotating biomedical entities and relations. Nucleic Acids Research, W390–W398 (2008)
[20] Tsai, R.T.-H., Wu, S.-H., Chou, W.-C., Lin, Y.-C., He, D., Hsiang, J., Sung, T.-Y., Hsu, W.-L.: Various criteria in the evaluation of biomedical named entity recognition. BMC Bioinformatics 7, 92 (2006)
[21] Saleem Basha, M.S., Dhavachelvan, P.: Web Service Based Secure E-Learning Management System - EWeMS. International Journal of Convergence Information Technology 5(7), 57–69 (2010)
[22] Hearst, M.A., Divoli, A., Guturu, H., Ksikes, A., Nakov, P., Wooldridge, M.A., Ye, J.: BioText Search Engine: beyond abstract search. ACM Portal, 2196–2197 (2007)

Exploring Possibilities of Reducing Maintenance Effort in Object Oriented Software by Minimizing Indirect Coupling

Nirmal Kumar Gupta and Mukesh Kumar Rohil

Birla Institute of Technology and Science, Pilani (India)
nirmalgupta@bits-pilani.ac.in,
rohil@bits-pilani.ac.in

Abstract. The quality of a good object oriented software design is much effective when it has highly maintainable class components. This paper describes an investigation into the use of indirect coupling to provide early indications of maintenance effort in object oriented software. The properties of interest are: (i) the potential maintainability of a class and (ii) the likelihood that a class will be affected by maintenance changes made to the overall system. The research explores that minimizing indirect coupling can provide useful indications of software maintenance effort that may have a significant influence on the effort during system maintenance and testing.

Keywords: Indirect Coupling, Software Maintenance Effort, Object Oriented Software, Software Quality.

1 Introduction

In Software Engineering, two design concepts (i.e. coupling and cohesion) are significant in developing good software. Coupling is the degree to which one component is dependent on other software components. A component with high coupling value is highly interdependent on other software components and vice versa. If a component is highly interdependent then any change in the component requires significant changes in other components to which it is coupled [1]. Hence highly coupled components require high maintenance effort. It can be noted that a system cannot completely be devoid of coupling for the proper functioning of a system. There is some need of some connections among various sub components (classes) of a system. Hence maintaining loose coupling among components is desirable characteristics of good software design [2].

Cohesion is a measure of how strongly related and focused are the responsibilities of a class. Strong cohesion has been recognized as a highly desirable property of classes [3]. If a subcomponent encapsulates unrelated or undesirable functionality in its shell then it means functionality has been poorly distributed among its subcomponents.

Object oriented software has various kinds of relationships among its components. Eder et al.[4] identify three different types of relationships. These relationships are interaction relationships between methods, component relationships between classes,

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 959–965.
springerlink.com

and inheritance between classes, are then used to derive different dimensions of coupling which are classified according to different strengths. Hitz and Montazeri [5] approach coupling by deriving two different types of coupling: object level coupling and class level coupling which are determined by the state of an object and the state of an object's implementation, respectively. While our understanding of coupling is improving, most research has been applied only to direct coupling which is, coupling between classes that have some direct relationship. However, there has been little investigation into indirect coupling, which is, coupling between classes that have no direct relationship. The discussion in existing literature just implies that indirect coupling is little more than the transitive closure of direct coupling.

Yang et al. [6] defines direct coupling as: a given class C is directly coupled to another class D, if there is an explicit reference to class D within the source code of class C. Class C has explicit reference to class D, if class D is the declared type of any expression or sub-expression contained within the source code of class C [6]. Direct coupling has compilation dependency which makes the dependent classes to undergo recompilation whenever a change is made in the class on which they are depending for proper functioning of the system. Therefore any change in coupled class D will requires subsequent change in class C also, otherwise class C may not compile. On the other hand, Indirect coupling which is transitive in nature is defined as: if a class X is coupled to class Y, which is in turn coupled to class Z, then class X is transitively dependent on class Z and hence indirectly coupled. Thus a modification on Z may cause a cascading effect along the connections, i.e. ripple effect [4] if there are more number of classes involved between X and Z in general. However, Indirect coupling manifests through hidden connections between seemingly unrelated parts of software system. This means, that this type of coupling exists but not visible through direct connection. Therefore it is very important to investigate this aspect of coupling and its impact over the maintainability of a class which contributes to the overall quality of the software.

Various researchers have worked in this area and tried to address the high coupling relating it with the maintenance effort by minimizing effect of coupling by keeping the value of indirect coupling as low as possible [5]. This indirect form of coupling has a stronger impact over maintenance effort as compared to direct coupling. This increased maintenance effort is due to the effort required in tracing and modification of the software components. It leads to increase in effort required with the increase in length and number of connections between software components [6]. Therefore one must try to achieve the value of indirect coupling as low as possible.

Measurement of indirect couplings can be achieved in many ways [6]. The most basic measure of coupling involves simply counting the number of other classes to which a given class has a linkage. If a STUDENT studies in a UNIVERSITY and admitted to some COURSE, and then assuming STUDENT, UNIVERSITY and COURSE are three classes, STUDENT would have a coupling value of 2. By this measure, an understanding of which classes are most coupled within the system can be made.

In this paper, we have identified the limitations of existing metrics in defining the maintenance effort in terms of indirect coupling existing in a class cluster within an object oriented software system. We also explore that indirect coupling minimization

defined by considering coupling paths we are able to relate indirect coupling to maintenance effort. This gives us a clear idea, how indirect coupling affects maintenance effort.

2 Related Work

Various researchers have put their efforts to define and measure various forms of couplings. According to Fenton and Pfleeger [7] "There are no standard measures of coupling". Many of the researches use some variation of Yourdon and Constantine's [8] original definition which defines as "a measure of the strength of interconnection" between modules. They suggests that coupling should be concretely defined in terms of the probability that coding, debugging, or modifying one module will necessary require consideration of changes of another module. Such kind of definitions is not formal definitions since they don't specify the meaning of "strength" or "interconnection". But such idea is an excellent heuristics for guiding the design of the software.

Berard [9] has given a comprehensive survey for object oriented coupling. He divides coupling into two basic categories: inter-face and internal coupling. Interface coupling exists when one object (client) makes use of another (server) object interface by calling its public method. In this case any change to the interface of server object enforces corresponding change in client object, but immune to any change occurs in the internals of classes of server object. Internal coupling occurs when an entity accesses an object's state, either from the "inside" or "outside". "Inside" Internal coupling occurs when that entity is a constituent of the accessed object, for example its method or a component object. "Outside" internal coupling occurs when that entity is a subclass or an unrelated object. He emphasizes that internal coupling is stronger and hence less desirable than interface coupling. Moreover, outside internal coupling is always stronger than its counterpart inside internal coupling.

Many researchers [1] [2] [8] have worked to understand object oriented coupling. In most of the cases the coupling metrics is described in terms of the features derived from source code itself. Such type of coupling is referred as direct coupling. Briand et al. [1] has done a survey and provided a framework which discusses various forms of coupling.

Eder et al. [4] gave another taxonomy of object-oriented coupling. He classifies coupling into three general forms: interaction, component and inheritance. Interaction coupling effectively refers to the following type of coupling: content, common, external, control, stamp and data coupling which are applied in the object oriented context, where the participants of coupling are methods and classes instead of modules. Component coupling concerns type usage relationship; class C is component coupled to C` if any of C's instance variable, local variables or method parameters of type C` are accessed by C. Component coupling represents compile time dependencies in the object oriented context. Finally, inheritance coupling refers to the inheritance relationship between a class and its direct or indirect subclass.

Chidamber and Kemerer [2] are first to define metrics Coupling Between Objects (CBO) for object-oriented coupling. They developed six design metrics for object oriented systems and analytically evaluated the metrics against Weyuker's [10] proposed set of measurement principles. The coupling for a class is the number of classes to which it is coupled. A class is deemed to be coupled to another class if it accesses

one or more of its variables or invokes at-least one of its methods. The inheritance based coupling is ignored. They state that high value of CBO means that high coupling value which result in high maintenance effort and therefore should be avoided.

Yang et al. [6] established that there is a form of coupling, which we call indirect coupling that has not been studied in depth and suggested that its existence may be the source of unnecessary maintenance costs. Poor software designs having large number of indirect coupling relations must be detected and ultimately such connections must be reduced.

3 Limitations

Based on our literature study we identify following limitations within the established metrics.

1. There are various issues with the definition of CBO. One is that definition is not specific as to whether a couple is counted in terms of instances or the number of classes.
2. Indirect coupling or strength between any two classes is measured as multiplications of direct coupling values along the path. The value of indirect coupling or strength leads to decrease as the path leads to increase since indirect coupling is multiple of direct coupling. The value of indirect coupling is maximum when the path length is one and leads to decrease as the path length increase. Consequently, it signifies that maintenance effort required decreases as path length increases and will be maximum when the path length is one.
3. The established metrics indirectly depends only upon the longest path even if the multiple paths exist between any two classes. If there are multiple paths exist between any two classes, then the value of indirect coupling is measured as maximum of various independent or shared path. Indirect coupling does not depend upon the number of paths existing between two classes, rather it only depends upon the path with highest indirect coupling value. So established metrics do not take into account the number of paths or number of connections existing between two classes.

4 Our Approach

In this section we will describe maintenance effort in object oriented software through indirect coupling. Indirect coupling can be understood with an analogy. Like in real life, you ask for your friend for particular task, which in turn asks to his friend so that the task which you have assigned to your friend could be completed. It means, there is direct relationship between you and your friend and also there is a direct relationship between your friend and his friend. So, there is no direct relationship between you and your friend's friend, but there is transitive relationship or indirect relationship between you and your friend's friend. In sense, there is effort in conveying the task to your friend's friend. Although, this relation seems to be hidden, but exists in real life. So, effort required to complete the task which you have assigned to your friend will be more than if it would had been done by your friend.

We consider the Indirect coupling defined by Yang et al. [6] as use-def relationships. These use-def relationships will extend from one class to some other class in a class cluster. In this particular form of indirect coupling we focus on its definition which is defined as "a given class C is indirectly coupled via data flow to another class D if and only if there exists a value used in class C that is defined in class D" [6]. The fundamental of this indirect coupling is that the behavior of class C is potentially dependent on the value generated by class D. Therefore applying any changes by modifying some value defined in class D may affect the behavior of class C. For example let us consider class definition in Fig. 1.

It is clear from Figure 1 that class A is directly dependent on classes B and C as it is creating their instances and calling their methods. If we try to rename classes B or C it will affect class A, as it would then require recompilation of A. Now indirect dependence is the complement of direct dependence. While indirect dependence may be thought of as just the transitive closure of direct dependences we find that this is not sufficient [11].

By same definition of 'dependent', we observe that C and D are not dependent. For example, removing D from the program would not cause compilation of C to fail. However a closer inspection reveals that a different kind of dependence exists between D and C. If we execute the program (through A's main method) results in a null pointer exception thrown by C because it tries to dereference the function aB.getString(), which evaluates to null. This is caused because D fails to initialize the value of the field str which is used by B through retVal() method. Now, if we uncomment the statement str = "Hi" in D we can avoid the null pointer exception in C. In other words there is a change to D that affects C, which signifies dependence, and which is an indirect dependency.

This type of dependency as described above w.r.t. Fig. 1 requires additional effort on the part of developer while performing any maintenance in the existing code.

```
class A
{
    public static void main(String[] args)
    {
        B aB = new B();
        aB.setString();
        C aC = new C();
        aC.readB(aB);
    }
}

class B
{
    String str;
    D aD;
    B()
    {
        aD = new D();
    }
    void setString()
    {
        str=aD.retVal();
    }
    String getString()
    {
        return str;
    }
}
```

```
class C
{
    void readB(B aB)
    {
        aB.getString().trim();
    }
}

class D
{
    String str;
    String retVal()
    {
        //str = "Hi";
        return str;
    }
}
```

Fig. 1. Interpreting Indirect Coupling

We use the concept of chains as introduced by Yang et al. [6] to define the metrics for such indirect coupling. A chain can be expressed in terms of graph vocabulary. Each statement in program would correspond to a node, while each immediate data flow from a definition site to a usage site corresponds to edges. We can define "length" of chains [6] based on the granularity level of measurement. This notion of distance can be mapped to maintenance effort, since the longer the chain of the flow of values across the system, the more work will be required to trace this flow, potentially having to switch between different methods and different classes or methods. The level of granularity can be determined by the level of boundary being considered, whether it is in terms of classes, methods or blocks. Selection of granularity of chains depends upon at what level we want to quantify the effort. In such case it is not a straight measure and depends upon the developer. For example using distance in terms of class boundaries is a simple measure, if we want to consider the coupling interactions between the classes. But this may be inappropriate if there are various self calls within same method and this effort cannot be taken into account at this level. To account for this one has to increase granularity at method level. Therefore a trade-off is required between maintenance effort and notion of chain length based on its granularity.

We argue that since the maintenance effort tends to increase as the path length increases between any two classes as one has to explore more number of classes. Therefore total effort will increase in an additive manner even though the value of indirect coupling will decrease. Similarly if multiple paths exist between any two classes, then the value of indirect coupling must consider all such existing paths instead of considering only path with highest indirect coupling value.

5 Conclusion and Further Research

In this paper we discussed use of a form of coupling known as indirect coupling, existence of which can be a source of additional (and often unnecessary) maintenance cost. Therefore its detection at early stage of development is very important. In this reference we have discussed the impact of indirect coupling over maintenance of the software. Whenever a class undergoes changes which may be because of a design defect or some enhancement in features which may result in modification of some existing class method, addition of some class method or even deletion of some class method, it may affect working of some other class in a class cluster. Identifying and removing such dependencies requires metrics for measurement, which can help to identify accurately such relationships.

References

1. Briand, L., Daly, W., Wust, J.: A Unified Framework for Coupling Measurement in Object-Oriented Systems. IEEE Transactions on Software Engineering 25, 91–121 (1999)
2. Chidamber, S.R., Kemerer, C.K.: Towards a Metrics Suite for Object Oriented Design. In: Proceedings of 6th ACM Conference on Object Oriented Programming, Systems, Languages and Applications (OOPSLA 1991), Phoenix, Arizona, pp. 197–211 (1991)
3. Dallal, J., Briand, L.: An object-oriented high-level design-based class cohesion metric. International Software Technology 52(12), 1346–1361 (2010)

4. Eder, J., Kappel, G., Schrefl, M.: Coupling and cohesion in object-oriented system, Technical report. Univ. of Klagenfurt (1994)
5. Hitz, H., Montazeri, B.: Measuring Coupling and Cohesion In Object-Oriented Systems. In: Proc. Int'l Symp. Applied Corporate Computing (ISACC 1995), Monterrey, Mexico, October 25-27 (1995)
6. Yang, H., Tempero, E.: Measuring the Strength of Indirect Coupling. In: Proceedings of the 2007 Australian Software Engineering Conference (ASWEC 2007), pp. 319–328. IEEE Computer Society, Washington, DC (2007)
7. Fenton, N.E., Pfleeger, S.L.: Software Metrics - A Rigorous & Practical Approach, ITP London (1997)
8. Yourdon, E., Constantine, L.: Structured Design: Fundamentals of a Discipline of Computer Program and System Design. Prentice-Hall (1979)
9. Berard, E.: Issues in the testing of object-oriented software. In: Electro 1994 International, pp. 211–219. IEEE Computer Society Press (1994)
10. Weyuker, E.: On testing non-testable programs. The Computer Journal 25(4), 465–470 (1982)
11. Yang, H.Y., Tempero, E.: Indirect Coupling As a Criteria for Modularity. In: Proceedings of the First International Workshop on Assessment of Contemporary Modularization Techniques (ACoM 2007), pp. 10–11. IEEE Computer Society, Washington, DC (2007)

A New Hashing Scheme to Overcome the Problem of Overloading of Articles in Usenet

Monika Saxena, Praneet Saurabh, and Bhupendra Verma

muskan.saxena@gmail.com, praneetsaurabh@gmail.com,
bk_verma3@rediffmail.com

Abstract. Usenet is a popular distributed messaging and files sharing service. Usenet flood articles over an overlay network to fully replicate articles across all servers. However, replication of Usenet's full content requires that each server pay the cost of receiving (and storing) over 1 Tbyte/day. This paper shows the design and implementation of Usenet database in Multilevel Hash table. A Usenet system that allows a set of cooperating sites to keep a shared, distributed copy of Usenet articles. In a standard multiple hashing scheme, each item is stored improves space utilization. This schemes open very amenable to Usenet implementation unfortunately this scheme occasionally require a large number of items to be moved to perform an insertion and deletion in Usenet database this paper shows that it is possible to significantly increase the space utilization of multiple choice hashing scheme by allowing at most one item to be moved during an insertion.

This paper represents the problems occur in this type of methods with little bit solution of them. Users may want to read, but it will not solve the problem of near exponential growth or the problems of Usenet's backbone peers.

Keywords: Usenet, Multiple Hashing scheme, Overloading of articles.

1 Introduction

The Usenet service has connected users world-wide. Users post articles into newsgroups which are propagated widely by an overlay network of servers. Users host lively discussions in newsgroups, because articles can represent multi-media files, cooperatively produce a large shared pool of files [1, 2]. A major attraction of Usenet is the incredible diversity and volume of content that is available [5].

Usenet is highly popular and continues to grow. Usenet provider upwards of 40,000 readers reading at an aggregate 20 Gbit/s several properties contribute to Usenet's popularity. Because Usenet's design aims to replicate all articles to all interested servers. Usenet user can publish highly popular content without the need to personally provide a server and bandwidth. Usenet's maturity also means that advanced user interfaces exist, optimized for reading threaded discussions or streamlining bulk downloads. However, Usenet service can be expensive, users post over 1 T byte/day of new content that must be replicated and stored [1, 5].

Usenet is the name of a worldwide network of servers for group communication between people. Usenet was created in 1979. It has seen an impressive growth from a

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 967–975.
springerlink.com

small academic community to a network used by millions of people from a wide variety of backgrounds all over the world [1]. The total size of the data flowing through Usenet has been more than tripling every year between 1993 and 2001[11,24]. This growth has not been without problems, and has raised significant challenges in how to handle the ever increasing volume of Usenet data flow. The amount of users and data they produce increases, with enough network bandwidth and storage capacity. Spending great sums of money on hardware components relieves the situation, but it does not solve the problem of database storage [2].

Selection of Hash Functions for Each Level

There are many hashing functions in literature for the purpose of reducing conflict keys and fast computation. Some of these functions perform very well in theory but in practice, their performance is very poor[10].

For example: A bucket size of 200. Experiment carried out with a test of data records of 500, 2,000, 5,000 and 10,000 on bucket size of 256 shows that hash values are either distributed at the beginning or in the centre or even at the end of the bucket. Figure shows the result of hash function of 500, 2,000, 5,000 and 10,000 records on 256 buckets. This leaves a lot of memory empty and rehashing of hash values into other tables is empty. [3]

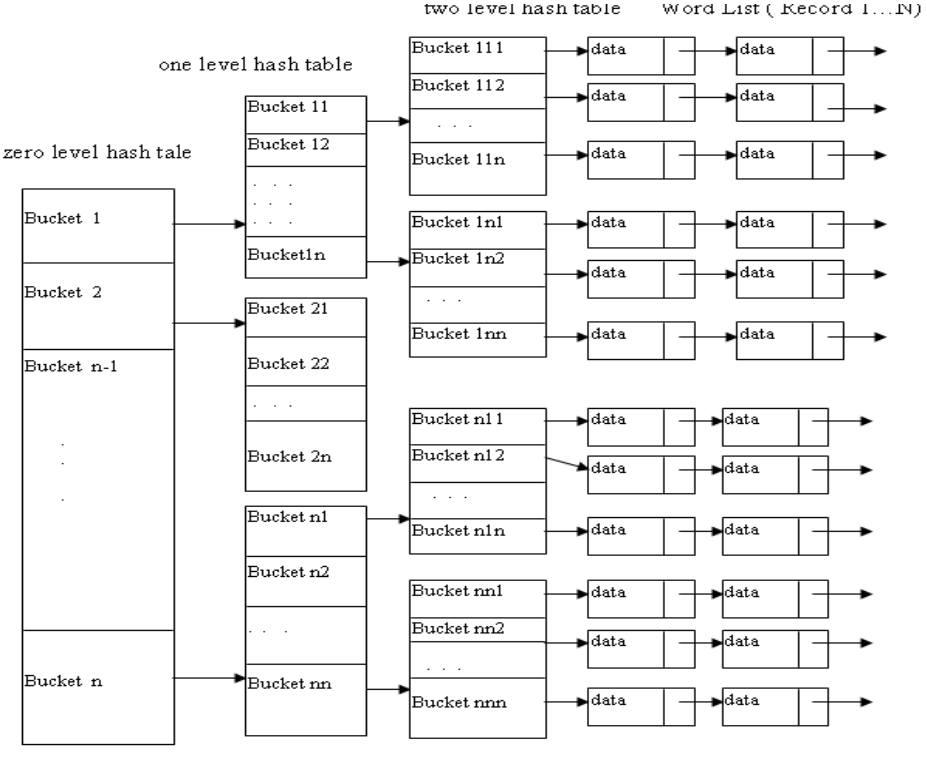


Fig. 1. Structure of Multilevel Hash Table Extension

The same experiment is carried out by using hashing codes derived by Robert J. Jenkins' with the same data records and the result shows that hash values are uniformly distributed. Figure 3 depicts Robert J. Jenkins hash codes of 500, 2,000, 5,000 and 10,000 records on 256 buckets. Robert J. Jenkin's hash codes place in every bucket of size 256 the records of 500 as 1 or 2, records of 2,000 as 7 or 8, records of 5,000 as 19 or 20 and records of 10,000 as 39 or 40. An attempt is made to test other discovered hash functions but none perfectly distributed hash values like Robert J. Jenkin's[23,29]. For example, Peter J. Weinberger and Arash Partow hash function do not perfectly hash values but uniformly distribute hash values across the buckets more than the size of 256. Three hashing codes namely, the hashing codes of Robert J. Jenkin, Peter J. Weinberger, and Arash Partow, were selected based on their performance by distributing hash values evenly in buckets of size more than 256[26,28].

2 Related Work

Usenet DHT [1] proposes a different procedure for Usenet system that allows a set of cooperating sites to keep a shared, distributed copy of Usenet articles. Usenet DHT consists of client-facing Usenet NNTP front-ends and a distributed hash table (DHT) that provides shared storage of articles across the wide area. This design allows participating sites to partition the storage burden, rather than replicating all Usenet articles at all sites.

News Cache [9] Proposes Caching techniques have not yet been adopted on a large scale. In this a high performance cache server for Usenet News that helps to conserve network bandwidth, computing power, and disk storage and is compatible with the current infrastructure and standards.

3 Problem Formulation

Multilevel hash tables must solve numerous problems in order to successfully store bulk data. For example, they must deal with problems:

1. Input/output Load
2. Traffic Control
3. Repetition of Articles.

The solution to these two classes is logically separate and invisible to applications. The application accesses the MHT using a simple key-value storage interface; this interface hides the complexity of the actual Distributed Hash Table implementation. The MHT implementation contains the logic for managing the constituent nodes, and then storing and maintaining the data. Nodes are organized by a routing layer that maintains routing tables and adapts them dynamically as nodes join and leave. The storage layer uses individual nodes as hash table buckets, storing a fraction of the application's data on each one.

Data maintenance, in the face of failures, is handled in the storage layer as well. The problem of node management and routing is well-studied. We rely on existing protocols such as Chord [1, 7] and Accordion [4] to handle this problem. This chapter

outlines the challenges faced by Multilevel Hash in providing durable data storage and the problems that Passing Tone must solve. We begin with a description of the requirements induced by multilevel hash table implementation for usenet database other applications with high request rate and high data volume.

4 Proposed System

Usenet targets mutually trusting organizations that can cooperate to share storage and network load. Prime examples of such organizations are universities, such as those on Internet2, that share high-bandwidth connectivity internally and whose commercial connectivity is more expensive. For such organizations, Usenet aims to:

1. Reduce bandwidth and storage costs in the common case for all participants;
2. Minimize disruption to users by preserving an NNTP interface; and
3. Preserve the economic model of Usenet, where clients pay for access to their local NNTP server and can publish content without the need to provide storage resources or be online for the content to be accessible.

In Proposed Solution,

Usenet Multilevel & Multiple Hashing Schemes

The scheme rarely moves an item in the Multilevel Hash Table, even though we have, in principle, allowed ourselves to perform at most one move per insertion operation, in many hardware applications, the frequency with which we perform moves may be much less important than the guarantee that we never perform more than one move per insertion. With this in mind, we introduce a scheme that is considerably more aggressive in performing moves while still guaranteeing that no insertion operation requires more than one.

Proposed Multilevel Hashing Scheme (Algorithm):

Step 1: for i=1 to d-1 do

Step 2: if Ti[hi(x)] is not full then

Step 3:Ti[hi(x)] <x

Step 4: Return y <- Ti[hi(x)]

Step 5: if Ti+1[hi+1(x)] is full then

Step 6: if Ti+1[hi+1(y)] is not full then

Step 7: Ti+1[hi+1(y)] <- y and Ti[hi(x)] <-x

Step 8: return if Td[hd(x)] is not full then

Step 9: Td[hd(x)] <- x

Step 10: else Add x to L

Step 11: Goto Step 2.

For intuition, consider inserting n items using the standard MHT insertion scheme. It is fairly clear, both from the definition of the scheme and from the differential equations, that as the items are inserted, the sub tables fill up from left to right, with newly inserted items cascading from Ti to Ti+1 with increasing frequency as fills up. Thus, it seems that a good way to reduce the overflow from the MHT is to slow down this cascade at every step. This idea is the basis for our new scheme, which we call the multilevel hashing scheme. The basic idea is that whenever an inserted item x cannot be placed in Ti, it checks whether it can be inserted into Ti+1. If it cannot be placed there, then rather than simply moving on Ti+2 to as in the standard scheme, the x item checks whether the item y in Ti[hi(x)]can be moved to Ti+1[hi+1(y)] . If this move is possible, then we move and replace it with. Thus, we effectively get a *multilevel hashing scheme* at preventing a cascade from Ti+1 to Ti+2[2].

This scheme is much more practical than it may first seem. To see this, consider a standard MHT implementation where we can read and hash one item from each sub table in parallel. In this setting, we can insert an item using the multilevel hashing scheme by reading and hashing all of items in T1[h1(x),…Td-1[hd-1(x)] in parallel. Once we have the hash values for these items, we can determine exactly how x should be placed in the table using the code.

5 Simulation and Results

Proposed solution is simulated with multilevel hashing scheme [2, 3], Our test deployment of Usenet with Multilevel Hash Table is able to support Usenet feed

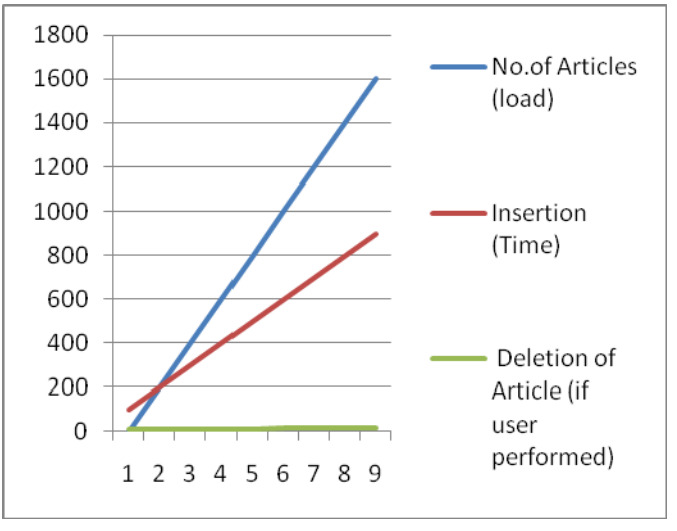
1. Input/output Load
2. Traffic Control
3. Repetition of Articles

Evaluation Method

The Multilevel hashing scheme is compared with Passing Tone scheme Distributed hash table [1] is evaluated under load using a live wide-area deployment. In this theory of research we compare the result of Usenet DHT with the result of multilevel hashing scheme using this deployment. The deployment consists of twelve machines at universities in the United States: four machines are located at MIT, two at NYU, and one each at the University of Massachusetts (Amherst), the University of Michigan, Carnegie Mellon University, the University of California (San Diego), and the University of Washington. Access to these machines was provided by colleagues at these institutions and by the RON test-bed. While these machines are deployed in the wide area, they are relatively well connected, many of them via Internet2. These machines are lightly loaded, stable and have high disk capacity. Each machine participating in the deployment has at least a single 2.4 Ghz CPU, 1 Gbyte of RAM and UDMA133 SATA disks with at least 120 Gbyte free. These machines are not directly under our control and are shared with other users; the write caching provided by the disks themselves is enabled on these machines. This may lead to higher write

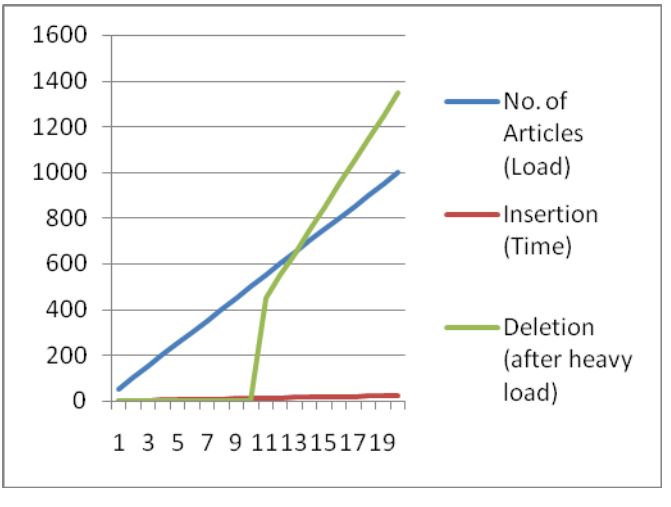
throughput, but synchronous writes requested by the DHash or UsenetDHT implementation (to ensure durability) may return without having actually been written to the disk.

Comparative Input/Output Load (Distributed and Multilevel Hash Tables)



Here: X-axis: Load or no. of articles & Y-axis: Time (ns).

Figure (a) Input/output Load in server using Distributed Hash Table using Passing Tone Scheme.



Here: X-axis: Load or no. of articles & Y-axis: Time (ns).

Figure (b) Input/output Load after using Multilevel Hash Table using multilevel hashing scheme.

Figure (a) shows the number of articles in given time on the server, the number of article shows the total load on server. The Usenet DHT not defines the concept of deletion of article in the server. So the load of server is comparatively very high. For example in figure (a) the article is inserted but not deleted so the **total load of the server is 1800.**

Figure (b) shows the number of articles in given time on the server, in multilevel hashing scheme the insertion of article in a single queue. As well as the multilevel hashing scheme defines the concept of deletion of article after one point of insertion. For example in figure (b) the deletion of article is started after the heavy load that is 400, it means the server contains only 400 articles when one new article is inserted the last one article is deleted from the server and stored in the other queue. The deleted article not present on server but when the user requires then the article calls from the queue of database. According to this the **total load on the server is 400** which is comparatively very less than the passing tone scheme.

6 Conclusion and Future Work

Usenet continues to be an important network service because of its distinct advantages over other data distribution systems. This results in over 1 Tbyte of new content posted to Usenet per day. Usenet servers have improved dramatically to carry this level of load, but the basic Usenet design hasn't changed, even though its flooding approach to distributing content is expensive. With the current design only a limited of servers can provide the full Usenet feed. We propose to exploit the recent advances in DHTs to reduce the costs of supporting .Three selected hash functions are used for multiple hash tables after testing their performance and ensure that they distribute hash values evenly in above 256 buckets. Each hash function is assigned to different-level of hash tables. The hash functions are written in Java Programming Language with time to look up for a given particular key. Data records of different sizes are employed to test the performance of the system.

References

[1] Sit, E., Morries, R., Frans Kaashoek, M., MIT CSAIL: Usenet DHT: A low-overhead design for usenet

[2] Kirsch, A., Mitzenmacher, M., Member IEEE: The power of one move hashing schemes for hardware

[3] Akinwale, A.T., Ibharalu, F.T.: The usefulness of multilevet hash tables with multiple hash functions in large databases

[4] Dabek, F., Kaashoek, M.F., Karger, D., Morris, R., Stoica, I.: Wide-area cooperative storage with CFS. In: Proc. of the 18th ACM Symposium on Operating Systems Principles (October 2001)

[5] Dabek, F., Sit, E., Li, J., Robertson, J., Kaashoek, M.F., Morris, R.: Designing a DHT for low latency and high throughput. In: Proc. of the 1st Symposium on Networked System Design and Implementation (March 2003)

[6] Ganger, G.R., Kaashoek, M.F.: Embedded inodes and explicit grouping: exploiting disk bandwidth for small files. In: Proc. of the 1997 USENIX Annual Technical Conference, pp. 1–17 (January 1997)
[7] Gradwell.com. Diablo statistics for news-peer.gradwell.net, http://news-peer.gradwell.net/ (accessed February 12, 2004)
[8] Grimm, B.: Diablo statistics for newsfeed. wirehub.nl (all feeders), http://informatie.wirehub.net/news/allfeeders/ (accessed February 12, 2004)
[9] Gschwind, T., Hauswirth, M.: NewsCache: A high performance cache implementation for Usenet news. In: Proc. of the 1999 USENIX Annual Technical Conference, pp. 213–224 (June 1999)
[10] Kantor, B., Lapsley, P.: Network news transfer protocol. RFC 977, Network Working Group (February 1986)
[11] Karger, D.R., Ruhl, M.: Diminished Chord: A Protocol for Heterogeneous Subgroup Formation in Peer-to-Peer Networks. In: Voelker, G.M., Shenker, S. (eds.) IPTPS 2004. LNCS, vol. 3279, pp. 288–297. Springer, Heidelberg (2005)
[12] Netwin. DNews: Unix/Windows Usenet news server software, http://netwinsite.com/dnews.htm (accessed November 9, 2003)
[13] Nixon, J., D'itri, M.: Cleanfeed: Spam filter for Usenet news servers, http://www.exit109.com/~jeremy/news/cleanfeed/ (accessed on February 15, 2004)
[14] Planet Lab, http://www.planet-lab.org
[15] Saito, Y., Mogul, J.C., Verghese, B.: A Usenet performance study (November 1998), http://www.research.digital.com/wrl/projects/newsbench/usenet.ps
[16] Stoica, I., Morris, R., Karger, D., Kaashoek, M.F., Balakrishnan, H.: Chord: A scalable peer-to-peer lookup service for Internet applications. In: Proc. of the ACM SIGCOMM (August 2001); An extended version appears in ACM/IEEE Trans. on Networking
[17] Zhao, B.Y., Huang, L., Stribling, J., Rhea, S.C., Joseph, A.D., Kubiatowicz, J.D.: A resilient globalscale overlay for service deployment. IEEE Journal on Selected Areas in Communications 22(1) (January 2004)
[18] Dabek, F., Zhao, B., Druschel, P., Kubiatowicz, J., Stoica, I.: Towards a Common API for Structured Peer-to-Peer Overlays. In: Kaashoek, M.F., Stoica, I. (eds.) IPTPS 2003. LNCS, vol. 2735, pp. 33–44. Springer, Heidelberg (2003)
[19] Decandia, G., Hastorun, D., Jampani, M., Kakulapati, G., Lakshman, A., Pilchin, A., Sivasubramanian, S., Vosshall, P., Vogels, W.: Dynamo: Amazon's highly available key-value store. In: Proc. of the 21st ACM Symposium on Operating System Principles (October 2007)
[20] Ford, B.: Structured streams: a new transport abstraction. In: Proc. of the 2007 ACM SIGCOMM (August 2007)
[21] Freedman, M.J., Freudenthal, E., Mazières, D.: Democratizing content publication with Coral. In: Proc. of the 1st Symposium on Networked Systems Design and Implementation (March 2004)
[22] Fritchie, S.L.: The cyclic news filesystem: Getting INN to do more with less. In: Proc. of the 9th LISA, pp. 99–112 (1997)
[23] Ghane, S.: Diablo statistics for spool-1t2.cs.clubint.net (Top1000 #24), http://usenet.clubint.net/spool-1t2/stats/ (accessed March 18, 2008)

[24] Ghane, S.: The official Top1000 Usenet servers page, http://www.top1000.org/ (accessed September 13, 2007)
[25] Ghemawat, S., Gobioff, H., Leung, S.-T.: The Google file system. In: Proc. of the 2003 19th ACM Symposium on Operating System Principles, Bolton Landing, NY (October 2003)
[26] Giganews. 1 billion usenet articles (April 2007), http://www.giganews.com/blog/2007/04/1-billion-usenet-articles.html
[27] Godfrey, P.B., Stoica, I.: Heterogeneity and load balance in distributed hash tables. In: Proc. of the 24th Conference of the IEEE Communications Society, Infocom (March 2005)
[28] Gradwell.com. Diablo statistics for news-peer.gradwell.net, http://news-peer.gradwell.net/ (accessed September 30, 2007)
[29] Gray, J., Mcjones, P., Blasgen, M., Lindsay, B., Lorie, R., Price, T., Putzolu, F., Traiger, I.: The recovery manager of the System R database manager. ACM Computing Surveys 13(2), 223–242 (1981)

Bio-inspired Computational Optimization of Speed in an Unplanned Traffic and Comparative Analysis Using Population Knowledge Base Factor

Prasun Ghosal[1], Arijit Chakraborty[2], and Sabyasachee Banerjee[2]

[1] Department of Information Technology, Bengal Engineering and Science University, Shibpur, Howrah 711110, WB, India
[2] Department of Computer Science and Engineering, Heritage Institute of Technology, Kolkata, WB, India
{prasung,arijitchakraborty.besu,sabyasachee.banerjee}@gmail.com

Abstract. Bio- inspired Computational Optimization of Speed in Unplanned Traffic and the comparative analysis is a very promising research problem. Searching for an efficient optimization method or technique to formulate optimal solution of a given problem in hand is very challenging and thereby to increase the traffic flow in an unplanned zone is a widely concerning issue. However, there has been a limited research effort on the optimization of the lane usage with speed optimization. This paper presents a novel technique to solve the problem optimally using the knowledge base analysis of speeds of vehicles, using partial modification of Bio Inspired Algorithm (Ant Colony Optimization) which, in turn will act as a guide and baseline for designing lanes optimally to provide better optimized traffic with less number of transitions between lanes.

1 Introduction

The hurdles in designing of non-accidental and non-congested lanes are required to move traffic safely and efficiently, although, highways and motor vehicles are designed to operate safely at speed. The purpose of this research is to create predictive models for different types of speed optimization techniques on lane, based on infrastructure design and traffic intensity. In this paper, the results for all transition points and vehicle's lane transition for speed optimization is discussed.

The Analysis starts with identifying basic issues and element of the problem in hand which are as follows.

- Entry zones,
- Transition points, and
- Exit zones.

Most of the traditional approach for handling the problem in hand is based on deterministic models which can be efficient and more or less accurate at times, but to achieve optimality of solution deterministically, at all time, seems to be far from reality till now.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 977–987.
springerlink.com

Non-deterministic approach can be used to tackle the inherent randomness of the problem, but at the same time it may not be accurate at all times.

Apart from that, making the lanes at their optimal average speed at any point of time using previous knowledge and current information is a major highlight presented in this paper.

This paper is divided into two sections. First section prevails the background of the present work with a description of the related works done so far in this area, and pointing out the drawbacks of the existing solutions. In the next section, the problem formulation and proposed algorithms are represented. First algorithm does not consider the concept of population knowledge base, and second one with the population knowledge base. Simulated results and observations are represented graphically in section 4. Finally, section 5 concludes the paper with possible future directions of work.

2 Background and Motivation

2.1 Related Works

The paper proposed by Jake Kononov, Barbara Bailey, and Bryan K. Allery, first explores the relationship between safety and congestion and then examines the relationship between safety and the number of lanes on urban freeways.

The relationship between safety and congestion on urban freeways was explored with the use of safety performance functions [SPF] calibrated for multilane freeways in Colorado, California, Texas.

The Focus of most SPF modeling efforts to date has been on the statistical technique and the underlying probability distributions. The modeling process was informed by the consideration of the traffic operations parameters described by the Highway Capacity Manual. [1]

H Ludvigsen, Danish Road Directorate, DK; J Mertner, COWI A/S, DK, 2006, published, Differentiated speed limits allowing higher speed at certain road sections whilst maintaining the safety standards are presently being applied in Denmark.

The typical odds that higher speed limits will increase the number of accidents must thus be beaten by the project.

The paper presented the methodology and findings of a project carried out by the Danish Road Directorate and COWI aimed at identifying potential sections where the speed limit could be increased from 80 km/h to 90 km/h without jeopardizing road safety and where only minor and cheaper measures are necessary. Thus it described how to systematically assess the road network when the speed limit is to be increased. [2].

C.J. Messer and D.B. Fambro, 1977, presented a new critical lane analysis as a guide for designing signalized intersections to serve rush-hour traffic demands.

Physical design and signalization alternatives are identified, and methods for evaluation are provided. The procedures used to convert traffic volume data for the design year into equivalent turning movement volumes are described, and all volumes are then converted into equivalent through-automobile volumes. [3].

Prasun Ghosal, Arijit Chakraborty, 2010 presented an idea of using lane buffers for arranging vehicles at their optimal speed in prefixed number of lanes. [4].

2.2 Drawbacks of Existing Solutions

Many traditional speed-optimizing algorithms for lanes were proposed earlier to optimize deterministic problems. But these algorithms didn't show their ability to use knowledge base to tackle the inherent randomness in the traffic systems. Therefore, to handle with such random realistic situation bio inspired algorithms are of great help and generate some efficient solution; good computational models of the same problem as well as good heuristics are required.

In order to put forward a feasible solution, we organize the problem in to two dimensions:

i) Designing an algorithm that will approximate about the no. Of lanes required.
ii) Placing the vehicle at an appropriate lane at any time t, so that all vehicles may move at their optimal speed and also developing a knowledge base.

3 Problem Formulations and Proposed Algorithms

3.1 Problem Description

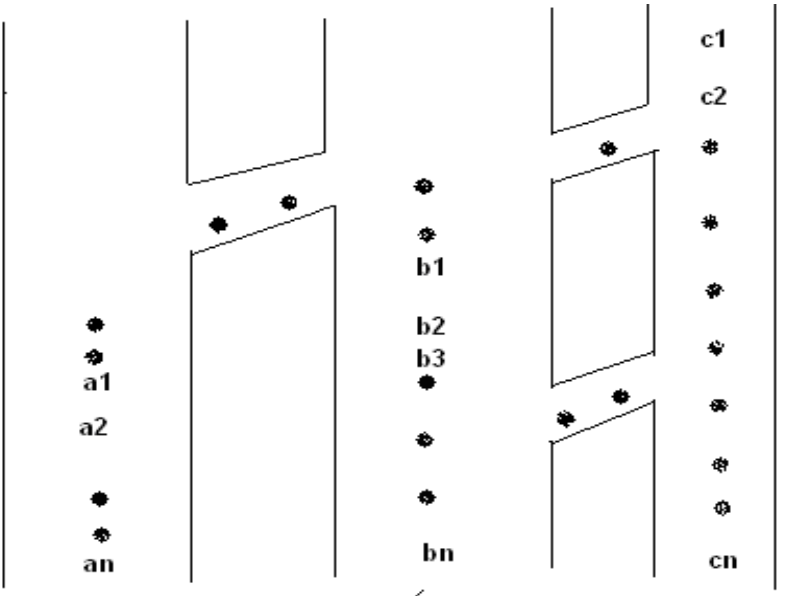


Fig. 1. Vertical lanes are unidirectional and with the property of the three lanes with transition points

Description of Figure 1

Figure 1**, T**hree vertical lanes that are unidirectional, and A = {a1, a2… an}, B = {b1, b2 ….bn}, C = {c1, c2,….,cn}, three lanes. I, II, III are the transition points through which vehicles can overtake its preceding vehicle with lesser speed and then immediately moves to its original lane. i.e. I from lane A to B or B to A and II, III are from B to C or C to B. Here we assume that each and every lane's car speed is greater than 0 kmph. If speed of any car is less than or equal to 0 kmph then we assume that there may be problem.

3.2 Problem Formulation

Random movement of vehicle in rush hour traffic are required to be frame up in optimal no. lanes with respect to number of transitions between lanes so that each lane have optimal speed.

Bio inspired algorithms like swarm intelligence technique used here with speed of the 'vehicle' acting as a pheromone to solve the problem in hand.

To maintain the optimality of a solution in a heuristic search using population information as a knowledge base is used in the proposed algorithms.

3.3 Proposed Algorithms

3.3.1 Algorithm I

Initial Assumptions

- There will be no change in the speed of the vehicle
- In case of sudden change of speed, accommodate the speed of previous slower vehicle.
- Any vehicle having speed equivalent of 0 is discarded from the initial sample or population; however, this is an assumption for the sake of simplicity.
- Any vehicle can overtake other vehicles (assumed for the sake of simplicity of the problem).

Details of the Proposed Algorithm

The major keynotes and functionality of the proposed algorithm are as follows: -

Step 1 is taking input from sensors, like the current speed of the vehicle, arrival time etc., and, counting the number vehicles the user has entered.

Step 2 is categorizing the vehicles depending on their current speed.

Step 3 is checking total how many numbers of lanes will be required for our sample data in an unplanned zone, and, which vehicle is moving in which lane.

Step 4 is checking total number of transitions i.e. at which point of the lane and from which lane to where the transition will occur.

Table 1. Symbolic Interpretation used in algorithms

Symbols used	**Meaning**
V_a	Velocity of vehicle a
V_b	Velocity of vehicle b
L_b	Lane of the vehicle b
Lc	Average speed of the lane C
type (i)	Category of Vehicle i
T	Arrival time difference between a high and low speed vehicles
t_a	Time interval to overtake a vehicles at lower speed
D	Distance covered by low speed Vehicle
d_a	Distance covered by high speeding Vehicle
B_n	Speed buffer of Lane n
Count	Total no. Vehicle in unplanned traffic
$Count_a$	Total no. Lanes for optimal speed
$Count_b$	Total number Of transition
Count_l	Population of a certain lane
$Count\_L_a$	Population of the a vehicle's lane

Pseudo Code (Algorithm I)

Input: Details of vehicles, Current speed of the vehicle, arrival time.
Output: Category of the vehicle, Number of lanes will be required, Number of transitions.
Step 1.1: Set count = 1; /*Used to count the number of vehicles. */
Step 1.2: take_ input (); /*Enter Details of vehicles, current speed, arrival time and store it into a record. */
Step 1.3: Continue Step 1.1 until sensor stops to give feedback and
Update count = count + 1 for each feedback;
Step 2: For 1 <= a <=count for each vehicle
If $0 < V_a < 11$ then categorize V_a as type A
If $10 < V_a < 31$ then categorize V_a as type B
If $30 < V_a < 46$ then categorize V_a as type C
If $45 < V_a < 51$ then categorize V_a as type D
If $50 < V_a < 101$ then categorize V_a as type E
Step 3: Set counter count1: = 1;
Set $L_1 = 1$;
For 2 <= a <= count for each Vehicle
For 1 <= b <= $count_a$
Compare the {type (a), type (b)} present in the lane

If different update $count_a = count_a + 1$ and
$L_a = count_a$;
Else
$L_a = b$;
End of loop;
End of loop;
Step 4: Set counter: $count_2 = count_1$;
For 1 <= a <= count -1 for each Vehicle
For 2 <= b <=count for each Vehicle
If type (a)= type (b) and $V_a < V_b$ and arrival time (V_a) <= arrival time (V_b)
Set t =arrivaltime (V_b) – arrivaltime (V_a);
Set $t_a = 0$;
Begin loop
Set $t_a = t_a + 1$;
Set $d = V_a * (t + t_a)$;
Set $d_1 = V_b * t_a$;
If d_1 <= d Set $count_b = count_b + 1$;
If $L_b = 1$ then transition will be to 2 - lane;
If $L_b = count_a$ then transition is $count_a$ - lane;
Else
Transition is either $L_b - 1$ or $L_b + 1$;
End loop;
End loop;
End loop;
Step 5: Return Number of lanes required = $count_a$;
Number of transitions required = $count_b$;
Step 6: End

Analysis of the Proposed Algorithm (Algorithm I)

- The above algorithm is implemented on an open unplanned Area.
- The objective will follow linear queue as long as speed/value/cost of proceeding is greater than the immediate next.
- Transition/Cross over are used and they again follow appropriate data structure in order to maintain the preceding step rule.
- Here we assume the lanes are narrow enough to limit the bi-directional approach.
- Here we maintain optimize speed for each lane.
- Here we also maintain the transition points if speed/value/cost of a vehicle is found unable to maintain the normal movement and transition in all the calculated lanes.
- Transition points are recorded with their position and number and it follows appropriate data structure in order to maintain the record.

3.3.2 Algorithm II

Description of the proposed algorithm. The primary sections of the proposed algorithm and their major functionalities are described below.

- Step 1. Take relevant information from sensors, i.e. the current speed of the vehicle, arrival time etc. and count the number of vehicles the sensor has entered along with that consider number of lanes that are present in the traffic.
- Step 2. Assign lanes to different vehicles having different current speeds at any time instant t in order to categorize them.
- Step 3. Determine whether the current speed of the vehicle is equal to the speeds present in speed buffers of lanes or not.
- Step 4. This step finds the lane, where, the difference between the vehicle's current speed and lane's speed buffer's average speed is minimum and takes the vehicle to the lane, categorizes it same as the lane's other vehicles, increases the population of the lane, and stores the vehicle's current speed in the speed buffer of the lane.
- Step 5. This step is used for checking total numbers of transitions, i.e. at which point of the lane and from which lane to where the transition will occur, thereby calculating the average speed of the lanes.

Pseudo Code (Algorithm II)

INPUT: Vehicle's name, current speed, arrival time.
OUTPUT: Vehicle's Type, Number of transitions.
Step 1.1: Set count=1; /*used to count the number of vehicles*/
Step 1.2: take_input ()/*Enter the inputs when speed of the vehicle is non-zero. */
Step 1.3: Continue Step 1.1 until sensor stops to give feedback.
Step 2: Set type (1)='A', Enter V1 into 1st lane's speed buffer, Set 1st lane's population (count_l) as '1', Set n=2.
For 2≤a≤count
Set a buffer buf=0
Loop1 until lane='0'
Loop2 for 1≤b<I for each vehicle
If Va = Vb
Set buf =1, type (a)=type (b)
Goto Step 3 and send Va to Step 3 as 'speed1'.
Step 2.1
If buf=1 then end Loop1
If buf=0
Enter Bn=Va, Set count_l=1, Set type(i)=A++;
End Loop1 /*Bn=n lane speed buffer*/
If lane=0
Then end Loop1.
If lane=0
Then end Loop.
Store buf2=i+1
Step 3: For 1≤a≤lane_l for each lane
If Bi's 1st speed=speed1
Update count_La++;
Set Bi, count_1=speed1
goto step 2.1

Step 4: for buf2≤a≤count
Set c=1, switch=0.
Set min=|Va, Lc|, /*Lc=c lane's average speed*/
type (i)=1st lane's vehicle type
For 1≤b≤lane_l
Set d=|Va, Lb|
If d=0
Set type (a)= type (Lb)
Update (b) lane's count_l= (b) lane's count_l+1
Set switch=1
End Loop
If d<min
Then min=d
Set type (a)= type (Lb)
Update (a) th lane's count_l= (b) lane's conut_l+1
Set (b) th lane's speed buffer [count_l] = (a) vehicle's speed (Va)
If switch=0
Update L1, count_1 ++;
Step 5: Set count2 as count2 =1
For 1≤a≤count-1
For 2≤b≤count
If type (a)= type (b) and Va<Vb and (a) vehicle's arrival time≤ (b) vehicle's arrival time
Set t=(a) vehicle's arrival time - (b) vehicle's arrival time
Set t1=0
Begin loop
Set t1=t1+1
Set d=Va*(t+t1)
Set d1=Vb*t1
If d1≤d set count2 = count2+1
If Lb =1 then transition will be to 2-lane
If Lb =count1 then transition will be to count1-lane
Else transition will be to Lb -1 or Lb +1
End loop
End loop
End loop
For 1≤m≤lane_l
Calculate each lane's average speed from its speed buffer.
Step 6: Return Number of transitions required= count2
Step 7: End.

Analysis of algorithm II: The salient points and features of the proposed algorithm may be analyzed as follows.

- The above algorithm is implemented on an open lane area.
- The objective will follow linear queue as long as speed/value/cost of proceeding to greater than the immediate next.

- Transition/Cross over are used and they again follow appropriate data structure in order to maintain the preceding step rule.
- Here we assume the lanes are narrow enough to limit the bidirectional approach.
- Here we also maintain the transition points if speed/value/cost of a vehicle is found unable to maintain the normal movement and transition in all the calculated lanes.
- Transition points are recorded with their position and number and it follows appropriate data structure in order to maintain the record.

4 Experimental Results and Observations

The optimization of the speed in rush hour traffic with the swarm intelligence approach in an open lane area used the population information as a knowledge base. Primary objective of this approach is to improve the traffic movement in rush hours and to optimize the speed of the vehicles using the concept of transition points between adjacent Lanes.

Proposed algorithms have been implemented with C++ in an open platform and executed using an Intel Pentium IV chip with 1GB memory.

Below is the simulated graphical analysis of experimental results, obtained thereby.

4.1 Simulated Graphical Analysis of the Proposed Algorithms

By implementing the above proposed algorithm and doing the simulation we were able to generate the following graphical results shown in figures 2 and 3 as follows.

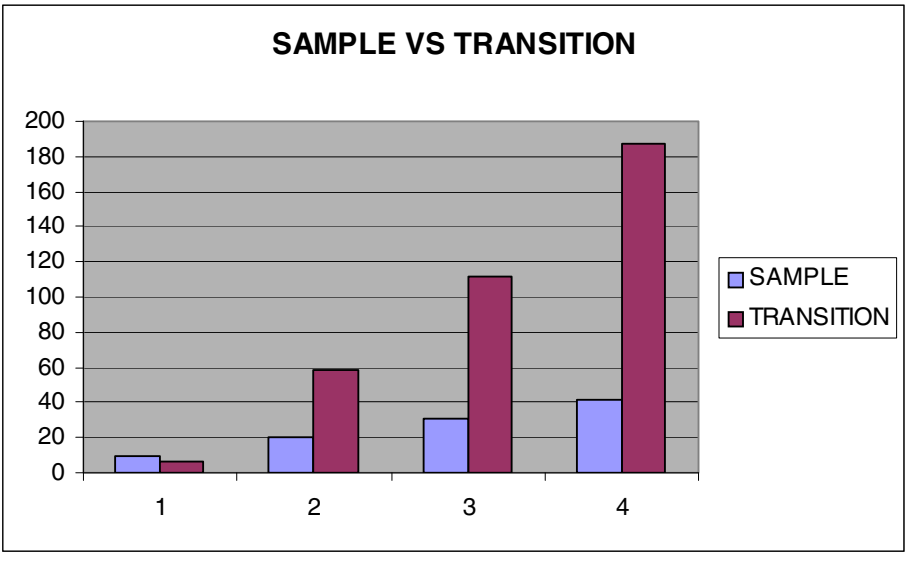


Fig. 2. This figure shows the variation of number of transitions with the number of lanes for a fixed number of samples i.e. 20

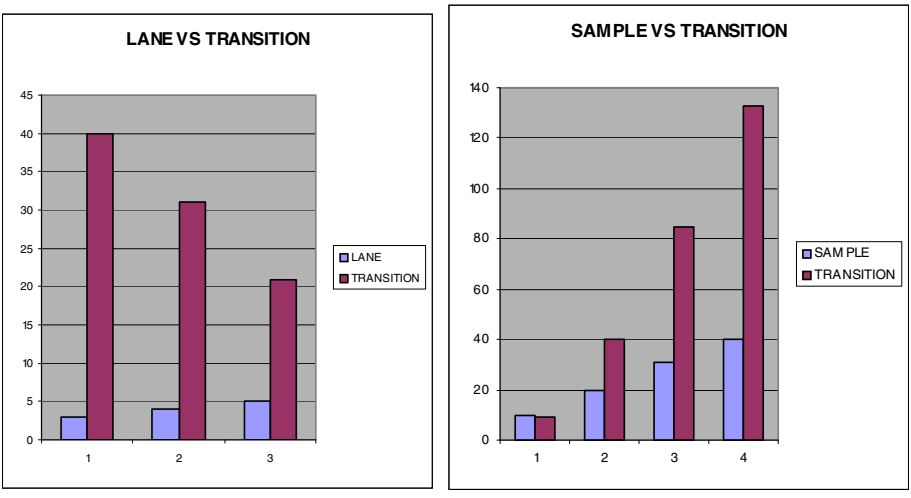


Fig. 3. First figure shows the nature of variation of the number of transitions with the variation of sample size without the population consideration. Second figure shows the same variation with population consideration.

Brief Analysis of Figure 2 and Figure 3

Analysis of the above-simulated results may be interpreted as follows.

- From figure 2 it is clear that as we increase the number of available lanes for a fixed number of samples, the number of transitions is decreasing drastically, which is, very much in conformity with the real life scenario.
- Another important point may be noticed from figure 3. As we are using the population knowledge base, there is a significant improvement in the number of transitions with the result when we were not using population knowledge base under consideration. This shows the effectiveness of our algorithm.

5 Conclusions and Future Scope

The article presented through this paper mainly emphasize on optimal usage of lanes using population information as knowledge base, but at the cost of transitions, because in real life scenario transitions may be too high, hence our future effort will be certainly in this direction.

In this article amount of time taken to transit between lanes has been considered as negligible. However cumulative sum of transition time between lanes in real world problem contributed much in optimality of the proposed solution.

Bio inspired algorithms (like swarm intelligence) has been used with population information as knowledge base, but partial modification of the stated concept taking weighted average of transition information as well as population information will certainly be taken into consideration during implementation and formulation of algorithms in future, there by optimizing various aspects of traffic movement in real world.

References

1. Kononov, J., Bailey, B., Allery, B.K.: The relationship between safety and congestion. Journal of the Transportation Research Board, No. 2083
2. Differentiated speed limits. In: European Transport Conference Differentiated Speed Limits (2007)
3. Messer, C.J., Fambro, D.B.: Critical lane analysis for intersection design. Transportation Research Record No. 644, 26–35 (1977)
4. Ghosal, P., Chakraborty, A., Das, A., Kim, T.-H., Bhattacharyya, D.: Design of Non-accidental Lane. In: Advances in Computational Intelligence, Man-Machine Systems and Cybernetics, pp. 188–192. WSEAS Press (2010)
5. Ghosal, P., Chakraborty, A., Banerjee, S.: Design of Efficient Knowledge Based Speed Optimization Algorithm in Unplanned Traffic. IUP Journal of Computer Sciences (in press)

Transliterated SVM Based Manipuri POS Tagging

Kishorjit Nongmeikapam[1], Lairenlakpam Nonglenjaoba[1], Asem Roshan[1], Tongbram Shenson Singh[1], Thokchom Naongo Singh[1], and Sivaji Bandyopadhyay[2]

[1] Dept. of Computer Science and Engg., Manipur Institute of Technology, Manipur University, Imphal, India
[2] Dept. of Computer Science and Engg., Jadavpur University, Jadavpur, Kolkata, India
{kishorjit.nongmeikapa,nonglen.ran,roshanasem99, tongbram.shenson,naongo.thokchom}@gmail.com, sivaji_cse_ju@yahoo.com

Abstract. Manipuri is a Scheduled Indian language which has two script: a borrowed Bengali Script and the original Meitei Mayek (Script). Manipuri is a resource poor language specially the Meitei Mayek text Manipuri. This paper deals with Support Vector Machine (SVM) based Part of Speech (POS) tagging of the Bengali Script text and then are transliterated to Meitei Mayek after POS tagging. So far POS tagging of Meitei Mayek Manipuri is not reported and this could be the first attempt.

Keywords: SVM, POS, Transliteration, Features, Manipuri.

1 Introduction

Part of Speech tagging is the task of labelling each word or token in a sentence with its appropriate syntactic category called part of speech. POS tagging has various applications in Natural Language Processing (NLP) systems like Information Retrieval, Summarization, Machine Translation, Name Entity Recognition (NER), Multiword Expression (MWE) identification, etc.

The paper is organized with related work in Section 2 followed by the concepts of SVM in Section 3, the transliteration algorithm in section 4, SVM tool and the feature selection in Section 5, the Model and the Experiment in Section 6 and the conclusion is drawn.

2 Related Works

Part of Speech taggers have been developed for several languages in this world. There are several works on POS taggers for English: a Simple Rule-based based POS tagger is reported in [1], transformation-based error-driven learning based POS tagger in [2], maximum entropy methods based POS tagger in [3] and Hidden Markov Model (HMM) based POS tagger in [4]. For Chinese, the works are found ranging from rule based, HMM to Genetic Algorithms [5]-[7]. For Indian languages like Bengali works are reported in [8]-[10] and for Hindi in [11]. Works of POS tagging using SVM methods can also be seen in [12]-[13].

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 989–999.
springerlink.com

Manipuri POS tagging is reported in [14]-[15] but so far POS tagging of Meitei Mayek Manipuri is not reported and this could be the first attempt. The identification of Reduplicated Multiword Expression (RMWE) is reported in [16]-[17]. Web Based Manipuri Corpus for Multiword NER and RMWEs Identification using SVM is reported in [18]. Transliteration work of Manipuri from Bengali Script to Meitei Mayek is reported in [19].

3 Concept of Support Vector Machine (SVM)

The idea of Support vector machines (SVM) were first shared by Vapnik [20]. In the work of [21] it is mention that Support Vector Machines is one of the new techniques for pattern classification which have been widely used in many application areas. The kernel parameters setting for SVM in training process impacts on the classification accuracy. Feature selection is another factor that impacts classification accuracy.

3.1 The Optimal Hyperplane (Linear SVM)

SVM concepts for typical two-class classification problems can be discussed for explanation. Given a training set of instance-label pairs $(x_i, y_i), i = 1,2, \ldots, m$ where $x_i \in R^n$ and $y_i \in_i \{+1, -1\}$, for the linearly separable case, the data points will be correctly classified by,

$$\langle w.x_i \rangle + b \geq +1 \; for \; y_i = +1 \tag{1}$$

$$\langle w.x_i \rangle + b \leq +1 \; for \; y_i = -1 \tag{2}$$

Combining Eqs. (1) and (2) into one set of inequalities.

$$y_i(\langle w.x_i \rangle + b) - 1 \geq 0 \; \forall \; i = 1, \ldots m \tag{3}$$

The SVM finds an optimal separating hyperplane with the maximum margin by solving the following optimization problem:

$$\text{Min}_{w,b} \frac{1}{2} w^T w \tag{4}$$

subject to: $y_i(\langle w.x_i \rangle + b) - 1 \geq 0$.

It is known that to solve this quadratic optimization problem one must find the saddle point of the Lagrange function:

$$L_p(w, b, \alpha) = \frac{1}{2} w^T.w - \sum_{i=1}^{m} (\alpha_i \, y_i (\langle w.x_i \rangle + b) - 1) \tag{5}$$

Where, the α_i denotes Lagrange multipliers, hence $\alpha_i \geq 0$. The search for an optimal saddle point is necessary because the Lp must be minimized with respect to the primal variables w and b and maximized with respect to the non-negative dual variable α_i. By differentiating with respect to w and b, the following equations are obtained:

$$\frac{\partial}{\partial w} L_p = 0, \; w = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{6}$$

$$\frac{\partial}{\partial w} L_p = 0, \; \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{7}$$

The Karush Kuhn–Tucker (KTT) conditions for the optimum constrained function are necessary and sufficient for a maximum of Eq. (5). The corresponding KKT complementarity conditions are:

$$\alpha_i[y_i(\langle w.x_i\rangle + b) - 1] = 0 \ \forall \ i \tag{8}$$

Substitute Eqs. (6) and (7) into Eq. (5), then L_p is transformed to the dual Lagrangian $L_D(\alpha)$,

$$\text{Max}_\alpha L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i.x_j\rangle \tag{9}$$

subject to $\alpha_i \geq 0, i = 1, \ldots, m$ and $\sum_{i=1}^{m} \alpha_i y_i = 0$.

To find the optimal hyperplane, a dual Lagrangian $L_D(\alpha)$must be maximized with respect to non-negative α_i. This is a standard quadratic optimization problem that can be solved by using some standard optimization programs. The solution α_i for the dual optimization problem determines the parameters $w^*$and b^* of the optimal hyperplane. Thus, we obtain an optimal decision hyperplane $f(x, \alpha^*, b^*)$ (Eq. (10)) and an indicator decision function sign $[f(x, \alpha^*, b^*)]$.

$$f(x, \alpha^*, b^*) = \sum_{i=1}^{m} y_i \alpha_i^* \langle x_i, x\rangle + b^* - \sum_{i \in sv}^{m} y_i \alpha_i^* \langle x_i.x\rangle \quad + b^* \tag{10}$$

In a typical classification task, only a small subset of the Lagrange multipliers α_i usually tends to be greater than zero. Geometrically, these vectors are the closest to the optimal hyperplane. The respective training vectors having nonzero α_i are called support vectors, as the optimal decision hyperplane $f(x, \alpha^*, b^*)$ depends on them exclusively.

3.2 The Optimal Hyper-Plane for Non-separable Data (Linear Generalized SVM)

The above concepts can also be extended to the non separable case, i.e. when Eq. (3) there is no solution. The goal is to construct a hyperplane that makes the smallest number of errors. To get a formal setting of this problem we introduce the non-negative slack variables $\xi_i \geq 0, i = 1, \ldots, m$. Such that

$$\langle w.x_i\rangle + b \geq +1 - \xi_i \ for \ y_i = +1 \tag{11}$$

$$\langle w.x_i\rangle + b \leq -1 + \xi_i \ for \ y_i = -1 \tag{12}$$

In terms of these slack variables, the problem of finding the hyperplane that provides the minimum number of training errors, i.e. to keep the constraint violation as small as possible, has the formal expression:

$$\text{Min}_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i \tag{13}$$

subject to : $y_i(\langle w.x_i\rangle + b) + \xi_i - 1 \geq 0, \xi_i \geq 0$.

This optimization model can be solved using the Lagrangian method, which is almost equivalent to the method for solving the optimization problem in the separable case. One must maximize the same dual variables Lagrangian $L_D(\alpha)$ (Eq. (14)) as in the separable case.

$$\text{Max}_\alpha L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i.x_j\rangle \tag{14}$$

subject to: $0 \leq \alpha_i \leq C, i, \ldots, m$ and $\sum_{i=1}^{m} \alpha_i y_i = 0$.

To find the optimal hyperplane, a dual Lagrangian $L_D(\alpha)$ must be maximized with respect to non-negative α_i under the constrains:

$\sum \alpha_i y_i = 0 \ and \ 0 \leq \alpha_i \leq C, i = 1, \ldots, m.$

The penalty parameter C, which is now the upper bound on α_i, is determined by the user. Finally, the optimal decision hyperplane is the same as Eq. (10).

3.3 Non-linear SVM

The nonlinear SVM maps the training samples from the input space into a higher-dimensional feature space via a mapping function Φ, which are also called kernel function. In the dual Lagrange (9), the inner products are replaced by the kernel function (15), and the non-linear SVM dual Lagrangian $L_D(\alpha)$ (Eq. (16)) is similar with that in the linear generalized case.

$$\left(\Phi(x_i).\Phi(x_j)\right) \coloneqq k(x_i x_j) \tag{15}$$

$$L_D(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k\langle x_i . x_j\rangle \tag{16}$$

subject to: $0 \le \alpha_i \le C, i = 1, \ldots, m$ and $\sum_{i=1}^{m} \alpha_i y_i = 0$.

This optimization model can be solved using the method for solving the optimization in the separable case. Therefore, the optimal hyperplane has the form Eq. (17). Depending upon the applied kernel, the bias b can be implicitly part of the kernel function. Therefore, if a bias term can be accommodated within the kernel function, the nonlinear SV classifier can be shown as Eq. (18).

$$\begin{aligned} f(x, \alpha^*, b^*) &= \sum_{i=1}^{m} y_i \alpha_i^* \langle \Phi(x_i).\Phi(x_j)\rangle + b^* \\ &= \sum_{i=1}^{m} y_i \alpha_i^* k(x_i, x) + b^* \end{aligned} \tag{17}$$

$$\begin{aligned} f(x, \alpha^*, b^*) &= \sum_{i \in sv} y_i \alpha_i^* \langle \Phi(x_i), \Phi(x_j)\rangle \\ &= \sum_{i \in sv} y_i \alpha_i^* k(x_i, x) \end{aligned} \tag{18}$$

Some kernel functions include polynomial, radial basis function (RBF) and sigmoid kernel, which are shown as functions (19), (20), and (21). In order to improve classification accuracy, these kernel parameters in the kernel functions should be properly set.

Polynomial kernel:

$$k(x_i, x_j) = (1 + x_i x_j)^d \tag{19}$$

Radial basis function kernel:

$$k(x_i, x_j) = exp\left(-\gamma \left\| x_i - x_j \right\|^2\right) \tag{20}$$

Sigmoid kernel:

$$k(x_i, x_j) = tanh(k x_i . x_j - \delta) \tag{21}$$

4 Transliteration Algorithm

The transliteration is the process of mapping a word of a source language script to another target language script. A simple transliteration scheme of Manipuri as in [19] is adopted here. Bengali which has 52 consonants and 12 vowels is mapped to Meitei Mayek which has 27 (Twenty seven) alphabets (Iyek Ipee) and its supplements: vowels, Cheitap Iyek, Cheising Iyek and Lonsum Iyek [22] are shown in Tables 1,2,3,4 and 5.

Table 1. Iyek Ipee characters in Meitei Mayek

Iyek Ipee			
ক->k (kok)	স(ছ,শ,ষ)->s (Sam)	ল->l (Lai)	ম->m (Mit)
প->p (Pa)	ন->n (Na)	চ->c (Chil)	ত(ট)->t (Til)
খ->S (Khou)	ঙ-> z (Ngou)	থ(ঠ)->H(Thou)	ব->w (Wai)
য(য়->y (Yang)	হ->h (Huk)	ঊ(ঊ)->U(Un)	ই(ঈ)->I(Ee)
ফ->f (Pham)	অ->A (Atia)	গ->g (Gok)	ঝ->J (Jham)
র->r (Rai)	ব->b (Ba)	জ-> j (Jil)	দ(ড)->d(Dil)
ঘ->G (Ghou)	ধ(ঢ)->D(Dhou)	ভ->v(Bham)	

Table 2. Vowels of Meitei Mayek

Vowel letters		
আ->Aa(Aa)	এ->Ae(Ae)	ঐ-AE(Ei)
ও->Ao(o)	ঔ->AE(Ou)	অং->Ax(aAng)

Table 3. Cheitap Iyek of Meitei Mayek

Cheitap Iyek			
ো->o (ot nap)	ি, ী-> i(inap)	া->a(aatap)	ে-> e(yetnap)
ৌ-> O (sounap)	ু, ূ-> u (unap)	ৈ-> E(cheinap)	ং-> x(nung)

Table 4. Cheising Iyek or numerical figures of Meitei Mayek

Cheising Iyek(Numeral figure)			
১->1(ama)	২->2(ani)	৩->3(ahum)	৪->4(mari)
৫->5(manga)	৬->6(taruk)	৭->7(taret)	৮->8(nipal)
৯->9(mapal)	১০->10(tara)		

Table 5. Lonsum Iyek of Meitei Mayek

Lonsum Iyek			
ক্-> K (kok lonsum)	ল্-> L (lai lonsum)	ম্->M (mit lonsum)	প্-> P(pa lonsum)
ণ্, ন্-> N (na lonsum)	ট্,ত্-> T (til lonsum)	ঙ্->Z(ngou lonsum)	ই, ঈ->I(ee lonsum)

Alphabets of Meitei Mayek are repeated uses of the same alphabet for different Bengali alphabet like ছ, শ, ষ, স in Bengali is transliterated to *s* in Meitei Mayek.

In Meitei Mayek, Lonsum Iyek (in Table 5) is used when ক্ is transliterated to *K,* ঙ্ transliterate to *Z*, ট্ transliterate to *T* etc. Apart from the above character set Meitei Mayek uses symbols like '>' (Cheikhie) for '।' (full stop in Bengali Script). For intonation we use '.' (Lum Iyek) and ' 'B (Apun Iyek) for *ligature*. Other symbols are as internationally accepted symbols.

Algorithm use for the transliteration scheme is as follows:

```
Algorithm: transliteration(line, BCC, MMArr[], BArr[])
1.    line : Bengali line read from document
2.    BCC : Total number of Bengali Character
3.    MMArr[] : Bengali Characters List array
4.    BArr[] : Meitei Mayek Character List array
5.    len : Length of line
6.    for m = 0 to len-1 do
7.     tline=line.substring(m,m+1)
8.     if tline equals blank space
9.       Write a white space in the output file
10.    end of if
11.    else
12.     for index=0 to BCC-1
13.      if tline equals BArr[index]
14.       pos = index
15.       break
16.      end of if
17.     end of for
18.     Write the String MMArr[pos] in the output file
19.    end of else
20.   end of for
```

In the algorithm two mapped file for Bengali Characters and corresponding Meitei Mayek Characters which are read and stored in the *BArr* and *MMArr* arrays respectively. A test file is used so that it can compare its *index* of mapping in the Bengali Characters List file which later on used to find the corresponding target transliterated Meitei Mayek Characters Combination. The transliterated Meitei Mayek Character Combination is stored on an output file.

5 SVM Tool and the Feature Selection

The idea of Support vector machines (SVM) are discussed in [20, 21]. Support Vector Machines is the new technique for pattern classification which has been widely used in many application areas. The kernel parameters setting for SVM in training process has an impact on the classification accuracy. Feature selection is another factor that impacts classification accuracy. A very careful selection of the feature is important in SVM. Various candidate features are listed. Those candidate features which are listed to run the system are as follows,

1. Surrounding words as feature: Preceeding word(s) or the successive word(s) are important in POS tagging because these words play an important role in determining the POS of the present word.

2. Surrounding Stem words as feature: The Stemming algorithm mentioned in [23] is used. The preceding and the following stemmed words of a particular word can be used as features. It is because the preceding and the following words influence the present word POS tagging.

3. Number of acceptable standard suffixes as feature: As mention in [23], Manipuri being an agglutinative language the suffixes plays an important in determining

the POS of a word. For every word the number of suffixes are identified during stemming and the number of suffixes is used as a feature.

4. Number of acceptable standard prefixes as feature: Same is the case for the prefixes. It also plays an important role for Manipuri language. For every word the number of prefixes are identified during stemming and the number of prefixes is used as a feature.

5. Acceptable suffixes present as feature: The standard 61 suffixes of Manipuri which are identified is used as one feature. As mention with an example in [23], suffixes are appended one after another. The maximum number of appended suffixes in Manipuri is reported as ten. So taking into account of such cases, for every word ten columns separated by a space are created for every suffix present in the word. A "0" notation is being used in those columns when the word consists of no acceptable suffixes.

6. Acceptable prefixes present as feature: 11 prefixes have been manually identified in Manipuri and the list of prefixes is used as one feature. For every word if the prefix is present then a column is created mentioning the prefix, otherwise the "0" notation is used. Upto three prefixes are considered for observation.

7. Length of the word: Length of the word is set to 1 if it is greater than 3 otherwise, it is set to 0. Very short words are generally pronouns and rarely proper nouns.

8. Word frequency: A range of frequency for words in the training corpus is set: those words with frequency <100 occurrences are set the value 0, those words which occurs >=100 are set to 1. The word frequency is considered as one feature since occurrence of determiners, conjunctions and pronouns are abundant.

9. Digit features: Quantity measurement, date and monetary values are generally digits. Thus the digit feature is an important feature. A binary notation of '1' is used if the word consist of a digit else '0'.

10. Symbol feature: Symbols like $,%, - etc. are meaningful in textual use, so the feature is set to 1 if it is found in the token, otherwise, it is set to 0. This helps to recognize SYM (Symbols) and QFNUM (Quantifier number) tags.

6 The Model and the Experiment

The model adopted here consists of two steps. At first, the SVM based POS tagger using the SVM rules mentioned in [20, 21], has been developed which performs classification by constructing an N dimensional hyperplane that optimally separates data into two categories. Running of the training process has been carried out by YamCha[1] toolkit, an SVM based tool for detecting classes in documents and formulating the POS tagging task as a sequential labelling problem. Here, the *pairwise* multi-class decision method and *polynomial kernel function* have been used. For classification, TinySVM-0.07[2] classifier is used.

In the second step, the output of the SVM based POS tagging in Bengali Script Manipuri text is used for transliteration to Meitei Mayek Manipuri.

6.1 Pre-processing and Feature Extraction for Running SVM

Running of the SVM based POS tagging sytem needs an input file of Bengali Script Manipuri text document. A separate input files are required both for training and

[1] http://chasen-org/~taku/software/yamcha/
[2] http://chasen-org/~taku/software/TinySVM/

testing. The training and test files consist of multiple tokens or words. In addition, each token consists of multiple (but fixed number) columns where the informations in the columns are used as a features. A sequence of tokens and other information in a line becomes a **sentence**. Before undergoing training and testing in the SVM the input document is converted into a multiple token file with fixed information column representing the values of the various. In the training file the last column is manually tagged with all the identified POS tags[3] whereas in the test file we can either use the same tagging for comparisons or only 'O' for all the tokens regardless of POS.

A total of 25,000 words are divided into two files, one consisting of 20000 words as training file and the second file consisting of 5000 words as testing file. The sentences are separated into equal numbers of columns representing the different features separated by blank spaces.

In order to evaluate the experiment result for POS tagging, the system used the parameters of Recall, Precision and F-score. These parameters are defined as follows:

$$\text{Recall, } \mathbf{R} = \frac{No\ of\ correct POS\ tag\ assigned by\ the\ system}{No\ of\ correct POS\ tag\ in\ the\ text}$$

$$\text{Precision, } \mathbf{P} = \frac{No\ of\ correct POS\ tag\ assigned by\ the\ system}{No\ of\ POS\ tag\ assigned by\ the\ system}$$

$$\text{F-score, } \mathbf{F} = \frac{(\beta^2 + 1)\ PR}{\beta^2 P + R}$$

Where β is one, precision and recall are given equal weight.

The experiment is performed with different combinations of features. The features are manually selected in such a way that the result shows an improvement in the F-measure. Among the different experiments with different combinations Table 7 lists some of the best combinations for POS tagging. Table 6 explains the notations used in Table 7.

Table 6. Meaning of the notations

Notation	Meaning
W[-*i*,+*j*]	Words spanning from the i^{th} left position to the j^{th} right position
SW[-*i*, +*j*]	Stem words spanning from the i^{th} left to the j^{th} right positions
P[*i*]	The *i* is the number of acceptable prefixes considered
S[*i*]	The *i* is the number of acceptable suffixes considered
L	Word length
F	Word frequency
NS	Number of acceptable suffixes
NP	Number of acceptable prefixes
D	Digit feature (0 or 1)
SF	Symbol feature (0 or 1)
DP[-i]	POS spanning from the previous i^{th} word's POS

[3] http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

Table 7. System performance with various feature combinations for POS tagging

Feature	R(in %)	P(in %)	FS(in %)
DP[1],W[-2,+1], SW[-1,+1], P[1], S[4], L, F, NS, NP, D, SF	**71.43**	**85.11**	**77.67**
DP[1],W[-2,+2], SW[-2,+1], P[1], S[4], L, F, NS, NP, D, SF	69.64	82.98	75.73
DP[2], W[-2,+3], SW[-2,+2], P[1], S[4], L, F, NS, NP, D, SF	67.86	80.85	73.79
DP[1], W[-3,+1], SW[-3,+1], P[1], S[4], L, F, NS, NP, D, SF	66.07	80.43	72.55
DP[3], W[-3,+3], SW[-3,+2], P[1], S[5], L, F, NS, NP, D	62.50	76.09	68.63
W[-3,+4], SW[-2,+3], P[2], S[5], L, F, NS, SF	50.00	64.37	56.28
W[-4,+1], SW[-4,+1], P[2], S[6], L, NP, D, SF	31.25	74.47	44.03
DP[3], W[-4,+3], SW[-3,+3], P[3], S[9], L, F, D, SF	30.00	66.67	41.38
W[-4,+4], SW[-4,+4], P[3], S[10], NS, NP	28.57	25.00	26.67

6.2 Evaluation and the Best Feature Set for POS Tagging

The best result for the SVM based POS tagging is the one which shows the best F-measure among the results. This happens with the following feature set:

F= { Dynamic POS tag of the previous word, W_{i-2}, W_{i-1}, W_i, W_{i+1}, SW_{i-1}, SW_i, SW_{i+1}, number of acceptable standard suffixes, number of acceptable standard prefixes, acceptable suffixes present in the word, acceptable prefixes present in the word, word length, word frequency, digit feature, symbol feature }

The experimental result of the above feature combination shows the best result which gives the Recall (**R**) of **71.43%**, Precision (**P**) of **83.11%** and F-measure (**F**) of **77.67%**.

6.3 Evaluation after Transliteration

The accuracy of transliterated output is measured in percentage by comparing the correctness in transliteration. The evaluation of the transliterated result shows an accuracy of **86.04%.** A lower score of **0.24%** is observed comparing with the claim in [19], which may be because of the domain of the corpus. The previous work was performed in the newspaper domain as it claim but here the experiment is performed in an article domain. Another factor of low accuracy is due to use of same character set of the Meitei Mayek relative to Bengali Script as mention in Section 4.

7 Conclusion

This work has its importance for the resource poor language like Manipuri. Collection of Meitei Mayek Manipuri text is a hard task thus transliteration is the only option to move on with the implementation of NLP applications for this language comparing to other popular Indian Language. This could be the first Meitei Mayek Manipuri POS tagging work using SVM. Works on the Meitei Mayek Manipuri text could be the future road map for research with new techniques of both POS tagging and transliteration.

References

1. Brill, E.: A Simple Rule-based Part of Speech Tagger. In: The Proceedings of Third International Conference on Applied NLP. ACL, Trento (1992)
2. Brill, E.: Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in POS Tagging. Computational Linguistics 21(4), 543–545 (1995)
3. Ratnaparakhi, A.: A maximum entropy Parts- of- Speech Tagger. In: The Proceedings EMNLP, vol. 1, pp. 133–142. ACL (1996)
4. Kupiec, R.: Part-of-speech tagging using a Hidden Markov Model. Computer Speech and Language 6(3), 225–242 (1992)
5. Lin, Y.C., Chiang, T.H., Su, K.Y.: Discrimination oriented probabilistic tagging. In: The Proceedings of ROCLING V, pp. 87–96 (1992)
6. Chang, C.H., Chen, C.D.: HMM-based Part-of-Speech Tagging for Chinese Corpora. In: The Proc. of the Workshop on Very Large Corpora, Columbus, Ohio, pp. 40–47 (1993)
7. Lua, K.T.: Part of Speech Tagging of Chinese Sentences Using Genetic Algorithm. In: The Proceedings of ICCC 1996, National University of Singapore, pp. 45–49 (1996)
8. Ekbal, A., Mondal, S., Bandyopadhyay, S.: POS Tagging using HMM and Rule-based Chunking. In: The Proceedings of SPSAL 2007, IJCAI, India, pp. 25–28 (2007)
9. Ekbal, A., Haque, R., Bandyopadhyay, S.: Bengali Part of Speech Tagging using Conditional Random Field. In: The Proceedings 7th SNLP, Thailand (2007)
10. Ekbal, A., Haque, R., Bandyopadhyay, S.: Maximum Entropy based Bengali Part of Speech Tagging. Advances in Natural Language Processing and Applications. Research in Computing Science (RCS) Journal (33), 67–78 (2008)
11. Singh, S., Gupta, K., Shrivastava, M., Bhattacharya, P.: Morphological Richness offsets Resource Demand–Experiences in constructing a POS tagger for Hindi. In: The Proceedings of COLING-ACL, Sydney, Australia (2006)
12. Antony, P.J., Mohan, S.P., Soman, K.P.: SVM Based Part of Speech Tagger for Malayalam. In: The Proc. of International Conference on Recent Trends in Information, Telecommunication and Computing (ITC), Kochi, Kerala, India, pp. 339–341 (2010)
13. Ekbal, A., Mondal, S., Bandyopadhyay, S.: Part of Speech Tagging in Bengali Using SVM. In: Proceedings of International Conference on Information Technology (ICIT), Bhubaneswar, India, pp. 106–111 (2008)
14. Doren Singh, T., Bandyopadhyay, S.: Morphology Driven Manipuri POS Tagger. In: The Proceeding of IJCNLP NLPLPL 2008, IIIT Hyderabad, pp. 91–97 (2008)
15. Doren Singh, T., Ekbal, A., Bandyopadhyay, S.: Manipuri POS tagging using CRF and SVM: A language independent approach. In: The Proceeding of 6th ICON 2008, Pune, India, pp. 240–245 (2008)
16. Kishorjit, N., Sivaji, B.: Identification of Reduplicated MWEs in Manipuri: A Rule based Approached. In: The Proc. of 23rd ICCPOL 2010, San Francisco, pp. 49–54 (2010)
17. Nongmeikapam, K., Laishram, D., Singh, N.B., Chanu, N.M., Bandyopadhyay, S.: Identification of Reduplicated Multiword Expressions Using CRF. In: Gelbukh, A.F. (ed.) CICLing 2011, Part I. LNCS, vol. 6608, pp. 41–51. Springer, Heidelberg (2011)
18. Doren Singh, T., Bandyopadhyay, S.: Web Based Manipuri Corpus for Multiword NER and Reduplicated MWEs Identification using SVM. In: The Proceedings of the 1st WSSANLP (COLING), Beijing, pp. 35–42 (2010)

19. Nongmeikapam, K., Singh, N.H., Thoudam, S., Bandyopadhyay, S.: Manipuri Transliteration from Bengali Script to Meitei Mayek: A Rule Based Approach. In: Singh, C., Singh Lehal, G., Sengupta, J., Sharma, D.V., Goyal, V. (eds.) ICISIL 2011. Communications in Computer and Information Science, vol. 139, pp. 195–198. Springer, Heidelberg (2011)
20. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer (1995)
21. Huang, C.-L., Wang, C.-J.: A GA-based feature selection and parameters optimization for support vector machines. Expert Systems with Applications 31, 231–240 (2006), doi:10.1016/j.eswa.2005.09.024
22. Mangang, K., Ng: Revival of a closed account. In: Sanamahi Laining Amasung Punsiron Khupham, Imphal, pp. 24–29 (2003)
23. Kishorjit, N., Bishworjit, S., Romina, M., Mayekleima Chanu, N., Bandyopadhyay, S.: A Light Weight Manipuri Stemmer. In: The Proceedings of National Conference on Indian Language Computing (NCILC), Chochin, India (2011)

A Survey on Web Service Discovery Approaches

Debajyoti Mukhopadhyay and Archana Chougule

Department of Information Technology
Maharashtra Institute of Technology
Pune 411038, India
{debajyoti.mukhopadhyay,chouguleab}@gmail.com

Abstract. Web services are playing an important role in e-business and e-commerce applications. As web service applications are interoperable and can work on any platform, large scale distributed systems can be developed easily using web services. Finding most suitable web service from vast collection of web services is very crucial for successful execution of applications. Traditional web service discovery approach is a keyword based search using UDDI. Various other approaches for discovering web services are also available. Some of the discovery approaches are syntax based while other are semantic based. Having system for service discovery which can work automatically is also the concern of service discovery approaches. As these approaches are different, one solution may be better than another depending on requirements. Selecting a specific service discovery system is a hard task. In this paper, we give an overview of different approaches for web service discovery described in literature. We present a survey of how these approaches differ from each other.

Keywords: WSDL, UDDI, indexing, service matching, ontology, LSI, QoS.

1 Introduction

Web services are application components which are based on XML [2]. Web services can be used by any application irrespective of platform in which it is developed. Web service description is provided in WSDL document. It can be accessed from internet using SOAP protocol. In industry, many applications are built by calling different web services available on internet. These applications are highly dependent on discovering correct and efficient web service. The discovered web service must match with the input, output, preconditions and effects specified by the user. Even after functional matching, QoS parameters also need to be matched to have best web service from available web services. Web services developed by different vendors are published on internet using UDDI[1]. UDDI is the mechanism for registering and discovering web services. It is platform independent registry as it is based on extensible markup language. It allows businesses to give list of services and describe how they interact with each other. In literature, many approaches for web service discovery are described some of which work on UDDI. Search in UDDI is based on keyword matching which is not efficient as huge number of web services may match a keyword and it is difficult to find the best one. Other approaches take advantage of semantic web concept where web

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 1001–1012.
springerlink.com

service matching is done using ontologies. Discovering web services automatically without human interface is an important concern. Different approaches to for automatic discovery of web services are also suggested by authors. This paper is organized as follows: section 2 gives overview of web service discovery process. Section 3 describes service discovery approaches. We conclude the paper in section 4.

2 Web Service Discovery

A web service discovery process is carried out in three major steps. First step is advertisement of web service by developers. Providers advertise web services in public repositories by registering their web services using web service description file written in WSDL [3]. Second step is web service request by user. User sends web service request specifying the requirement in predefined format to web service repository. Web service matcher which is core part of web service discovery model, matches user request with available web services and finds a set of web service candidates. Final step is selection and invocation of one of the retrieved web services. Discovery of correct web service depends on how mature web service matching process is. i.e.; how actual requirements of user are represented in formalized way and how they are matched with available services.

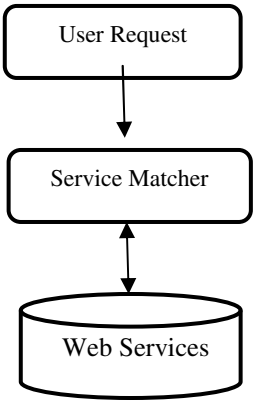


Fig. 1. Web service discovery

3 Survey

In this section we give overview of thirteen different approaches for web service discovery. For each one, we mention the details where one approach differs from others.

3.1 Context Aware Web Service Discovery

As format for sending web service request is fixed, some information in user's request is lost during transforming user's request to formalized one. To overcome this limitation, context aware web service discovery approach is suggested by Wenge Rong and Kecheng Liu [4]. Context aware discovery is useful for request

optimization, result optimization and personalization. As concept of context is very complex, they suggest with an example that context should be domain oriented or problem oriented. The context in web service discovery is formally defined as any information that explicitly and implicitly affects the user's web service request generation. They divide context in two categories as Explicit and implicit. Explicit context is directly provided by the user during matchmaking process such as Q&A information. Implicit context is collected in automatic or semi-automatic manner. Implicit context is more applicable to web service discovery as user is not directly involved. Context awareness is again divided in four categories depending on how context is collected. The categories are Personal profile oriented context, Usage history oriented context, Process oriented context and other context. Personal profile oriented context is collected using user's personal profile which contains personal data, preferences and other information. Personalization information such as location, time and user's situation is used for decomposing the discovery goal, setting selection criteria and supplying parameters. Limitation of this method is, it makes system architecture more complicated when new attributes and constraints are introduced. Usage history oriented context is collected for predicting user's next behavior. It is based on assumption that web service requests by specific user are similar during a certain period of time. Usage history oriented context is again divided in two categories as Personal usage history oriented context and Group usage history oriented context. User's previous system interaction can be stored in system log. Log records can be used to provide recommendation for service selection decision. But the user may not have similar requirements afterwards. So Group usage history oriented context is used where web service matchmaking is based on behavior information of other user groups in similar situation. One of the examples of Group oriented context awareness is collaborative filtering (CF) which may be memory based or content based. Group oriented context can also be collected from observation data in particular community. Process oriented context is built from user sessions. User reactions to retrieved web services are understood in particular request session and the discovery is optimized. This feed based process oriented context can be built using Probabilistic Latent Semantic Analysis (PLSA) [5]. In case where single web service is not sufficient to complete user request, composition of multiple web services is carried out. In this case, context should be built considering composite web service discovery process. This approach is better than traditional keyword matching used in UDDI as user intension is understood better.

3.2 Publish Subscribe Model

Falak Nawz, Kamram Qadir and H. Farooq Ahmad[6] propose push model for web service discovery where service requesters are provided with service notification prior to discovery. They use semantic based web service matching where service descriptions are matched using OWL-S [7], an ontology language for web service description. They also rank published web services depending on the scores assigned using concept matching. They divide the system in two phases as subscription phase, which starts when a subscriber registers himself onto registry for notification of required services and notification phase, which starts when a new service is published on registry. In subscription phase, when user goes for subscribing, subscription information along with his/her location and specific web service requirements are

stored in subscription knowledge base. Information from knowledge base is used later for service matching. Information in knowledge base is stored in OWL format. Service categories are maintained according to user requests received till date. In each service category, lists are maintained containing information about number of input and output parameters required by each subscription. The best matching web service is selected by matching user requirements (inputs, outputs, preconditions and effects) to OWL-S descriptions stored in registry. Matching can be in one of six levels as Exact, Plug-In, Subsume, Enclosure, Unknown and Fail. In notification phase, OWL-S descriptions for newly registered web services are added to matching subscription categories and listing for number of parameters is updated. If there is no single matching subscription category, then service descriptions are directly added to registry. Subscriptions in knowledge base are stored in the form of subsumption relationship ontology. Subscribers are notified when their leasing subscription time expires so that subscribers can renew their subscriptions. Time required for web service discovery is minimized with this approach as search area is reduced to specific category. Probability of finding most suitable web service also increases. Limitation of this approach is, it adds overhead in developing and maintaining new components in system architecture.

3.3 Keyword Clustering

Web service discovery based on Keyword clustering and concept expansion is suggested by J. Zhou, T. Zhang, H. Meng, L. Xiao, G. Chen and D. Li[8]. They calculate similarity matrix of words in domain ontology based on Pareto principal and use that for semantic reasoning to find matching service. Bipartite graphs are used to find matching degree between service requests and available services. They describe in detail how Kuhn-munkres algorithm can be used to compute optimal matching of a bipartite graph.

3.4 Service Request Expansion

One more approach for enhancing web service discovery is sending modifying user requests as suggested by A. Paliwal, N. Adam and C. Bornhovd [9]. They expand service requests by combining ontologies and latent semantic indexing. They build the service request vector according to the domain ontology, build the training set of the LSI classifier by extracting features from selected WSDL files, and then project the description vectors and the request vector. They utilize the cosine measure to determine similarities and to retrieve relevant WSDL service descriptions. Ontology linking is done using semi automated approach. It is done by mapping domain ontologies to upper merged and mid-level ontologies. Keywords are selected from service request by pre-processing service request which includes removal of mark-ups, punctuations and use of white spaces etc. Keyword based search is applied to upper ontology and relevant ontology is identified from ontology framework. Service request is expanded by acquiring associated concepts related to initial service request with semantic matching and assembling of concepts and enhanced service request is achieved. From collection of WSDL documents, relevant WSDL documents are found and service description set is built. Service description set is then transformed

into a term-document matrix by parsing and processing of documents. Removal of mark-ups and index entries, removal of punctuations, stoplist, use of white space as term delimiters and stemming to strip word endings are the steps involved in WSDL document processing. Term-document matrix is generated out of WSDL processing which indicates term frequencies. Built training set is used for LSI. LSI includes Singular Value Decomposition (SVD). In SVD, original matrix is approximated by a linear combination of a decomposition set of term to text-object association data. The resulting description vectors and request vectors are then projected and similarity is calculated using cosine similarity. At last, resulting web services are ranked based on similarity measure. Disadvantage of this approach is cost of computing LSI and string SVD is high.

3.5 BPEL Processes Ranking Using Graph Matching

When user requests for web service in available web services repository if exact matching web service does not exist, then approximate matching web service can be suggested by service matcher. To achieve this goal, behavioral matching is required. D. Grigori, J. Carlos Corrales, M. Bouzeghoub and A. Gate [10] developed matching technique which works on BPEL [11] behaviour model. User requirements are expressed as a service behaviour model. They transform BPEL specification to a behaviour graph using flattening strategy and transform service matching problem to graph matching problem. In graphs, regular nodes represent the activities and connectors represent split and join rules. Flattening strategy maps structural activities to respective BPEL graph fragments. The algorithm traverses the nested structure of BPEL control flow (BCF) in a top-down manner and applies recursively a transformation procedure specific to each type of structured activity. This procedure checks whether the current activity serves as target and source for links and adds arcs or respective join and split connectors in the resulting graph fragment. Five structural activities handled are Sequence, Flow, Switch, While and Pick. The generated process graph which represents user requirements is then compared with the target graphs in library. Error correcting graph matching is used to find approximate matching process model if exact matching process graph is not available. Similarity is measured as inverse of distance between two graphs representing BPEL. Distance is defined as cost of transformations needed to adapt the target graph in order to cover a subgraph in the request graph. Different measures for calculating cost are defined as distance between two basic BPEL activities, matching links between connector nodes and linguistic similarity between two labels based on their names. The results are optimized by applying granularity-level analyzer. It checks whether composition/decomposition operations are necessary for graph matching. BPEL processes are then ranked in decreasing order of calculated distance between graphs and web services. The limitation of this approach is method is completely based on syntactic matching. Semantics of user request is not considered.

3.6 Layer Based Semantic Web Service Discovery

Finding a matching web service in whole service repository is time consuming process. Guo Wen-yue, Qu Hai-cheng and Chen Hong [12] have divided search in three layers by applying filters at each layer and thus minimizing search area. They

have applied this approach on intelligent automotive manufacturing system. Three layers for service matching are service category matching, service functionality matching and quality of service matching. Semantic web service discovery is done based on OWL-S, using ServiceProfile documents for service matching. First step in service discovery is service category matching. Service category matching is carried out to minimize time and storage space required for service matching. At this layer, service category matching degree is computed. ServiceCatogory attribute in ServiceProfile contains category of service. This value is matched against service category of request which is passed b user while sending request. If there is match, web service is selected to enter the next service functionality matching layer. Advertisements that do not meet the demands are filtered out. Then service functionality matching degree is computed in the service functionality matching layer. For functionality matching, four attributes defined in ServiceProfile are matched against service request. These attributes are hasInput, hasOutput, hasPrecondition and hasResult. Advertisements that do not meet the conditions are filtered out, while other advertisements that satisfy the conditions are selected to enter the next quality of service matching layer. Last step is computing quality of service matching degree. Quality of service is decided based on response time of service discovery and reliability of service discovery system. From service category matching degree, service functionality matching degree and quality of service matching degree, service matching degree is calculated and the advertisements that best meet needs of requesters are presented to requesters in the form of list.

3.7 Service Discovery in Heterogeneous Networks

Web services are heavily used by military networks which are heterogeneous and decentralized in nature. There is need of interoperable service discovery mechanism to enable web service based applications in military networks. Traditional mechanisms for web service discovery such as UDDI and ebXML are not suitable in military networks as they are centralized and cannot be available during network partitioning.. F. Johnsen, T. Hafsoe, A. Eggen, C. Griwodz and P. Halvorsen[13] suggest the web service discovery solution which can fulfil the requirements in military networks. As same protocol cannot be used in heterogeneous networks, they suggest using of service discovery gateways, so that each network domain can employ the most suitable protocol. Interoperability is ensured by using service discovery gateways between the domains that can translate between the different service discovery mechanisms. Creation and interpretation of service descriptions in clients, servers, and gateways are done to ensure interoperability. This mechanism is called as Service Advertisements in MANETs (SAM), a fully decentralized application-level solution for web services discovery. It integrates periodic service advertisements, caching, location information, piggybacking, and compression in order to be resource efficient. A gateway periodically queries all services in the WS-Discovery and proprietary domains. Services that are available (if any) must then be looked up in the gateway's local service cache. This local cache is used to distinguish between services that have been discovered, converted, and published before, and new services that have recently appeared in each domain. If a service is already present in the cache, it has been converted and published before, and nothing needs to be done. On the other hand, if the service is not in the cache, it is

translated from one service description to the other, published in the network, and added to the local cache. For each query iteration, gateway compares local cache containing all previously found services with the list of services found now. Service is removed from the local cache if it deleted from its domain. This behaviour allows the gateway to mirror active services from one domain to the other, and remove any outdated information. They have implemented a gateway prototype solving transparent interoperability between WS-Discovery and a cross-layer solution, and also between WS-Discovery and SAM.

3.8 Web Service Indexing

To enable fast discovery of web services, available web services can be indexed using one of the indexing mechanisms such as inverted indexing and latent semantic indexing. B. Zhou, T. Huan, J. Liu and MeizhouShen [14] describe how inverted indexing can be used for quick, accurate and efficient web service discovery. In semantic web service discovery, user request is matched against OWL-S descriptions of web services. In this case, inverted index can be used to check whether the OWL-S description with the given id contains the term. Inverted index consists of list of keywords and frequency of keyword in all OWL-S documents. Every keyword is connected to a list of document ids in which that keyword occurs. They have suggested extensions to inverted lists to find positions of terms in OWL-S descriptions.

M. Aiello, C. Platzer, F. Rosenberg, H. Tran, M. Vasko and S. Dustdar[15] describe VitaLab system which is web service discovery system based on indexing using hashtable. They have implemented indexing on WSDL descriptions which are parsed using Streaming API for XML (StAX). Two hash tables namely parameter index and service index are built. Parameter table maintains the mapping from each message into two lists of service names for request and response respectively, to get a list of services that consume or produce a particular message. Service index maps service names to their corresponding detail descriptions. Generated indexes are serialized as binary files and stored in non-volatile memory and used the same every time when new service is added or existing service is modified or deleted.

One more index structure for concept based web service discovery is used by C. Wu, E. Chang and A. Aitken[16]. They use Vector Space Model (VSM) [17] indexes and Latent Semantic Analysis (LSA) indexer on term document matrices generated by processing WSDL descriptions which are retrieved by web crawlers. Term document matrices are generated according to Zipf Law and by applying Singular Value Decomposition (SVD). VSM indexer takes all term documents as input and outputs WSDL indices representing term document matrices. These term matrices are given as input to LSA indexer which generates as output semantic space for service retrieval.

Advantage of indexing approach is, once the indexes are available, it is easy to retrieve the objects fast using index. Limitation of the approach is, indexing process is computationally expensive. It requires additional space to store the indexes and the indexes need constant update if data changes often.

3.9 Structural Case Based Reasoning

Georgios Meditskos and Nick Bassiliades[18] describe semantic web service discovery framework using OWL-S. They detail a web service matchmaking algorithm which extends object-based matching techniques used in Structural Case-based Reasoning. It allows retrieval of web services not only based on subsumption relationships, but also using the structural information of OWL ontologies. Structural case based reasoning done on web service profiles provide classification of web services, which allows domain dependent discovery. Service matchmaking is performed on Profile instances which are represented as objects considering domain ontologies. In Semantic case based reasoning (SCBR), similarity is measured as interclass similarity considering hierarchical relationships and intraclass similarities by comparing attribute values of objects of same class. Web service discovery is done by measuring similarity at three levels as taxonomical similarity, functional similarity and non-functional similarity. Taxonomical similarity between advertisements and query is the similarity of their taxonomical categorization in a Profile subclass hierarchy. It is calculated using DLH metric which represents the similarity of two ontology concepts in terms of their hierarchical position. Four hierarchical filters for matching are defined as exact, plugin, subsume and sibling. Functional similarity is calculated based on input and output similarity (signature matching) of advertisement and query. It is ensured whether all the advertisement inputs are satisfied by query input and all query outputs are satisfied by advertisement outputs. Non-functional similarity is measured by directly comparing values of data types and objects. They calculate overall similarity between advertisement and query in terms of their taxonomical, functional and non-functional similarity. The semantic web service discovery framework is further enhanced to perform service discovery using ontology roles as annotation constraints. The framework is implemented using OWLS-SLR [19] and compared with OWLS-MX matchmaker.

3.10 Agent Based Discovery Considering QoS

As there can be multiple web services available providing same kind of functionality, best service among them should be selected. This can be done using QoS parameters. T. Rajendran and P. Balasubramanie[20] suggest a web service discovery framework consisting of separate agent for ranking web services based on QoS certificates achieved from service publishers. Main entity of web service discovery framework is verifier and certifier which verifies and certifies QoS of published web service. The service publisher component is responsible for registration, updating and deletion of web service related information in UDDI. Service publisher is supplied with business specific and performance specific QoS property values of web services by service providers. Verification and certification of these properties is then done by web service discovery agent. After that, service provider publishes its service functionality to UDDI registry through service publisher. The service consumer searches UDDI registry for a specific service through discovery agent which helps to find best quality service from available services which satisfies QoS constraints and preferences of requesters. QoS verification is the process of validating the correctness of information described in service interface. Before binding the web service, service consumer

verifies the advertised QoS through the discovery agent. The result of verification is used as input for certification process. Backup of certificates is also stored by web service agent which is used for future requests for similar kind of web services. Time required for selecting web service with best QoS values eventually decreases. The QoS parameters for selecting best web service are suggested by authors. These parameters are response time, availability, throughput and time. Values of these parameters are stored in tModels of respective web services which are supplied by service publisher.

3.11 Collaborative Tagging System

In feedback based web service discovery, comments from users who already have used the web services can be useful for other users. This approach is adapted by U. Chukmol, A. Benharkat and Y. Amghar[21]. They propose collaborative tagging system for web service discovery. Tags are labels that a user can associate to a specific web service. Web services are tagged by different keywords provided by different users. For each tag, tag weight is assigned. A tag weight is the count of number of occurrences of a specific tag associated to a web service. Tag collection is the collection of all entered tags. Each tag in tag collection is associated to a certain number of web services, forming resource vector. They employ both types of tags as keyword tag and free text tags. These tags are made visible to all users who want to access web services. When user sends a query for keyword based discovery, matching web service is found by checking whether there exists a web service having tag matching exactly with user input. If not, it checks whether synonym to the keyword exists with help of synonym set obtained from WordNet. System provides support for preparing queries using AND, OR and NOT operation. User can also attach more than one keyword as tag. This is called as free text tagging. In this case, each web service is associated with multiple keyword tags. Keywords are arranged one below another called as aggregated text tag. It is then transformed in vector of terms. Service discovery using free text is also provided. When user sends query as free text, it is converted into vector of terms to find matching web service. Vector Space Model is employed to carry out vector term matching. They use the Porter stemming algorithm [22] to extract terms vector from a query document and the aggregated text tag associated to web service. Similarity is calculated as is the cosine value between the two vectors representing both texts. Resulted web services are ranked according to the values of cosine coefficient between the query text and all aggregated text tags associated to resources.

3.12 Peer-to-Peer Discovery Using Finite Automaton

As centralised web service discovery approach has many disadvantages such as single point of failure and high maintenance cost, F. Emekci, O. Sahin, D. Agrawal and A. Abbadi[23] propose a peer-to-peer framework for web service discovery which is based on process behaviour. Framework considers how service functionality is served. All available web services are represented using finite automaton. Each web service is defined as follows: A Web service p is a triple, p=(I, S, R), such that, I is the implementation of p represented as a finite automaton, S is the service finite

automaton, and R is the set of request finite automata. When user wants to search for web service, PFA of finite automaton of web service(R) is sent for matching. Matching is done against S by hashing the finite automata onto a Chord ring. Chord is a peer-to- peer system for routing a query on hops using distributed hash table. Regular expression of the queried PFA is used as the key to route the query to the peer responsible for that PFA.

3.13 Hybrid Approach

Main categories of web service discovery approaches are keyword based and ontology based. Y. TSAI, San-Yih, HWANG and Y. TANG [24] make use of both approaches for finding matching web service. The approach considers service providers information, service descriptions by providers, service description by users, operation description by providers, tags and categories and also QoS attributes. For finding similarity between query and candidate web service, similarity between two operations is calculated first. For this, similarity between input/output of query and web service is calculated using ontology of web service. For a given operation input (output), each of its message part is mapped to a concept in an ontology. This similarity is tested at three levels. First relative positions of the concepts associated to query and web service message parts are considered. The position can be one of three parts, 1) exact where two classes are same, 2) subsume where one concept is super class of other class and 3) others where two concepts are not related. At second layer, similarity is measured based on paths between two concepts and at last similarity is measured based on information content (IC) of concepts. After considering similarity between input/output attributes, other attributes such as name, description, tags and operation name are considered. For these attributes, text based method is used where two words are compared lexically measuring the longest continuous characters in common. For measuring overall similarity, weights are assigned to all the attributes using analytical hierarchy process. AHP is the method for multiple criteria decision making [25]. Overall similarity between query and web service is calculated as weighted sum of similarities of all associated attributes. Described approach is tested and compared with text based approach and ontology based approach and it is shown that, hybrid approach gives better results than using each approach separately.

4 Conclusion

Success of published web services depends on how it is getting discovered. Efficiency, accuracy and security factors must be considered while providing discovery mechanism. We have given overview of different web service discovery approaches with their advantages and disadvantages. Many approaches differ in the way web service matching is carried out. Some approaches are considering concept of semantic web, while some other focus on information retrieval methods. Some approaches suggest enhancement in web service request based on metadata about web services generated by feedback of other users. Some approaches suggest additional tools in traditional framework of web service discovery. Minimizing total search area using clustering techniques is also suggested. Survey shows that considering QoS

parameters while selecting is important because, number of available web services providing same kind of functionality is very large. As web service discovery requiring manual interference may take more time, solutions for automatic discovery are drawing more attention.

References

1. UDDI. org., UDDI Spec Technical Committee Draft, http://www.uddi.org/pubs/uddi_v3.htm#_Toc85907967
2. Web Services Architecture, http://www.w3.org/TR/ws-arch/
3. Christensen, E., Curbera, F., Meredith, G., Weerawarana, S.: Web Services Description Language (WSDL) 1.1, W3C Note (2001)
4. Rong, W., Liu, K.: A Survey of Context Aware Web Service Discovery: From User's Perspective. In: Fifth IEEE International Symposium on Service Oriented System Engineering (2010)
5. Hofmann, T.: Probabilistc Latent Semantic Analysis, Stockholm. Uncertainity in Artificial Intelligence, UAI (1999)
6. Nawaz, F., Qadir, K., Farooq Ahmad, H.: SEMREG-Pro: A Semantic based Registry for Proactive Web Service Discovery using Publish Subscribe Model. In: Fourth International Conference on Semantics, Knowledge and Grid. IEEE Xplore (2008)
7. w3. org., OWL-S: Semantic Markup for Web Services, http://www.w3.org/Submission/OWL-S/
8. Zhou, J., Zhang, T., Meng, H., Xiao, L., Chen, G., Li, D.: Web Service Discovery based on Keyword clustering and ontology
9. Paliwal, A.V., Adam, N.R., Bornhovd, C.: Web Service Discovery: Adding Semantics through Service Request Expansion and Latent Semantic Indexing. In: International Conference on Services Computing (SCC 2007). IEEEXplore (2007)
10. Grigori, D., Corrales, J.C., Bouzeghoub, M., Gater, A.: Ranking BPEL Processes for Service Discovery. IEEE Transactions on Services Computing 3(3) (July-September 2010)
11. oasis-open.org, Web Services Business Process Execution Language Version 2.0, http://docs.oasis-open.org/wsbpel/2.0/wsbpel-v2.0.pdf
12. Guo, W.-Y., Qu, H.-C., Chen, H.: Semantic web service discovery algorithm and its application on the intelligent automotive manufacturing system. In: International Conference on Information Management and Engineering. IEEEXplore (2010)
13. Johnsen, F., Hafsøe, T., Eggen, A., Griwodz, C., Halvorsen, P.: Web Services Discovery across Heterogeneous Military Networks. IEEE Communications Magazine (October 2010)
14. Zhou, B., Huan, T., Liu, J., Shen, M.: Using Inverted Indexing to Semantic WEB Service Discovery Search Model. In: 5th International Conference on Wireless Communications, Networking and Mobile Computing. IEEEXplore (2009)
15. Aiello, M., Platzer, C., Rosenberg, F., Tran, H., Vasko, M., Dustdar, S.: Web Service Indexing for Efficient Retrieval and Composition. In: Proceedings of the 8th IEEE International Conference on E-Commerce Technology and the 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (2006)
16. Wu, C., Chang, E., Aitken, A.: An empirical approach for semantic Web services discovery. In: 19th Australian Conference on Software Engineering. IEEEXplore (2008)

17. Wong, S.K., Ziarko, W., Wong, P.C.: Generalized vector spaces model in information retrieval. In: The Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 1985, Montreal, Quebec, Canada, June 05-07 (1985)
18. Meditskos, G., Bassiliades, N.: Structural and Role-Oriented Web Service Discovery with Taxonomies in OWL-S. IEEE Transactions on Knowledge and Data Engineering 22(2) (February 2010)
19. OWLS-SLR (2008), http://lpis.csd.auth.gr/systems/OWLS-SLR
20. Rajendran, T., Balasubramanie, P.: An Optimal Agent-Based Architecture for Dynamic Web Service Discovery with QoS. In: Second International Conference on Computing, Communication and Networking Technologies. IEEEXplore (2010)
21. Chukmol, U., Benharkat, A., Amghar, Y.: Enhancing Web Service Discovery by using Collaborative Tagging System. In: 4th International Conference on Next Generation Web Services Practices. IEEEXplore (2008)
22. The porter stemming algorithm, http://snowball.tartarus.org/algorithms/porter/stemmer.html
23. Emekci, F., Sahin, O.D., Agrawal, D., El Abbadi, A.: A Peer-to-Peer Framework for Web Service Discovery with Ranking. In: Proceedings of the IEEE International Conference on Web Services, ICWS 2004 (2004)
24. Tsai, Y.-H., Hwang, S.-Y., Tang, Y.: A Hybrid Approach to Automatic Web Services Discovery. In: International Joint Conference on Service Sciences. IEEEXplore (2011)
25. Saaty, T.L.: Decision making with the analytic hierarchy process. International Journal of Services Sciences 1(1), 83–98 (2008)

Intensity Based Adaptive Fuzzy Image Coding Method: *IBAFC*

Deepak Gambhir and Navin Rajpal

[1] Research Scholar, University School of IT, Guru Gobind Singh Inderprastha University, Dwarka, New Delhi
and
Assistant Professor, Amity School of Engineering and Technology, Bijwasan New Delhi
gambhir.deepak@gmail.com
[2] Professor & Dean, University School of IT, Guru Gobind Singh Inderprastha University, Dwarka, New Delhi
navin_rajpal@yahoo.com

Abstract. A new design method of image compression as Intensity Based Adaptive Fuzzy Coding (*IBAFC*) is presented. In this design, the image is decomposed to non overlapping square blocks and hence each block is classified as either edge or smooth blocks. This classification based upon some predefined Threshold compared to adaptive quantization level of each block. Then each block is coded as either fuzzy F-transform compressed for edge block or mean value of block is sent for smooth block. The experimental results proves that the proposed *IBAFC* scheme is superior to conventional AQC and Intensity based AQC (IBAQC) on measures like MSE PSNR alongwith visual quality.

Keywords: Image Coding, Fuzzy Logic, Quantization.

1 Introduction

Image compression is a process of removing redundant data in an image matrix, such as to reduce storage costs and transmission time. Compression algorithms are always under wide attention because of its capability to reduce the represent able image data to the order of more than 20, thus there would be hard to see any difference between original image compared to decompressed image. Hence the removal of redundant information avaliable in image matrix, the methods of image compression/decompression are concerned with compressing image, to minimum bit per pixel possible and reproducing it to nearly perfect to original image reconstruction.

Edges in the image are the efficient representation of complete image, edges thus it is quite useful to have edge based image compression systems. A very early work on edge and mean based image compression presented in [10], is an example of static image compression for low bit rate applications.

An edge preserving lossy image coding is presented in [5] by villegas et al. Here the edge image is obtained from the original image using four different edge detection methods: Canny, Sobel, Roberts and Prewitt operators, then the original image is

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 1013–1022.
springerlink.com

wavelet or contourlet transformed, and a pixel mapping is performed based on wavelet domain image. For the compression, the selected edges points and the approximation image (which determines the compression factor) are selected.

An another interesting compression algorithm Intensity based adaptive quantization coding (IBAQC) method by Azawi et al based on the conventional adaptive quantization coding (AQC) is presented in [7]. In this method, the image is divided to blocks and the image blocks are further classified into edge or non edge blocks according to a predefined set threshold. For an edge block, the complete block data is coded according to AQC, whereas for a non edge block, only min value of the block is coded.

There is another interesting method similar to adaptive quantization coding algorithm is block truncation coding (BTC) avaliable in literatue and to the date, many variants are available in literature for BTC and its improvements. BTC is improved according to some error diffusion in [4] whereas the BTC for color image is improved in [3].

A recent image compression method using fuzzy complement edge operator utilizing the basic Block Truncation Coding (BTC) algorithm [8] is developed by Amarunnishad et al. This method is based on replacement of the conventional BTC block with the fuzzy logical bit block (LLB) such that the sample mean and standard deviation in each image block are preserved. This fuzzy logical bit block is obtained from the fuzzy edge image by using the fuzzy complement edge operator (YIFCEO) based on their past paper [9] of fuzzy edge detection. The input image is encoded with the block mean and standard deviation like BTC and the fuzzy logical bits.

An another recent method of fuzzy image compression/decompression based on fuzzy F-transform (FTR) introduced in [2] by F.D.Martino et. al. In this method, the image is first decomposed to normalized blocks and then these normalized blocks are fuzzy compressed to lower dimensions image sub-matrices by using Gaussian types membership functions. Then on the decompression side, the lower dimension sub-image is expanded to their original size and then de-normalized. In its simulation results, the authors claimed that, this compression method results similar to JPEG for compression rate and good PSNR.

Another Fuzzy Gradient based adaptive lossy predictive coding system for gray scale images are presented in [1] by El Khamy et. al. This image coding method, employs the adaptive fuzzy prediction methodology in the predictor design and in addition to this, author also claims that this system adopts a novel fuzzy gradient-adaptive quantization scheme and this system also possesses superior performance over their non-fuzzy counterparts. This is possible because of the adaptivity in the fuzzy prediction methodology and as well as the gradient-adaptive quantization scheme.

The rest of paper proceeds as follows: in Section 2 there is detailed discussion about three different image compression methods 1) fuzzy F-transform 2) AQC and 3) IBAQC. Section 3 details about proposed IBAFC algorithm and in Section 4 the results and discussions are presented and Section 5 concludes the paper.

2 Compression Methods

The two past but effective method of Image Compression is described as:

2.1 Fuzzy F-Transform Based Image Coding Algorithm

The fuzzy based F-transform (FTR) [6] is used to compute the fuzzy logical bit plane for an image compression F-transform based system. This is an extension of 1-D framework to 2-D frame for F-transform based compression. To compute the fuzzy bit plane there is need to define fuzzy partition set $[A_1\ A_2\ A_3 \ldots A_n]$ between the closed interval $[a,b]$ having number of uniform points in between as $p_1, p_2, p_3 \ldots p_n$ such that $a = p_1 < p_2 < p_3 < \ldots < p_n = b$. Hence this fuzzy partition $[A_1\ A_2\ A_3 \ldots A_n] : [a,b] \rightarrow [0,1]$ is said to hold following conditions.

1. $A(p_i) = 1$ for $\forall i = 1,2,3 \ldots n$.
2. $A_i(p)$ is a continous function for $\forall i = 1,2,3 \ldots n$.
3. $A_i(p)$ is strictly increasing on $p_i - 1\ to\ p_i$ and $A_i(p)$ is strictly decreasing on $p_i\ to\ p_i + 1$.
4. For all p belongs to $[a,b], \sum A_i(p) = 1$.
5. The fuzzy Partition $[A(p_1)\ A(p_2)\ A(p_3) \ldots\ A(p_n)]$ is said to be uniform when $p_i - (p_i - 1) = (p_i + 1) - p_i$. i.e for $n \geq 3$ & $p_i = a + h(i-1)$ where $h = (b-a)/(n-1)$ for $\forall i = 1,2,3 \ldots n$.
6. $A_i(p_i - p) = A_i(p_i + p)$ for every $p \in [0,h]$.
7. $A_i + p = A_i(p - h)$ for every $x \in [x_i, x_i + 1]$.

Thus for example, the fuzzy partition functions are:

On the basis of this fuzzy partition the one dimensional continuous and discrete fuzzy F transform variable is defined as:

$$F_x = \frac{\int_a^b f(P_y)A_x(P_y)dP_y}{\int_a^b A_x(P_y)dP_y} \qquad F_x = \frac{\sum_{y=1}^{m} f(P_y)A_x(P_y)}{\sum_{y=1}^{m} A_y(P_y)} \tag{1}$$

Hence this framework of 1-D can be extended to the 2-D Image Fuzzy F-transform as:

$$F_{i,j} = \frac{\sum_{x=1}^{n}\sum_{y=1}^{m} I(x,y)A_i(x)B_j(y)}{\sum_{x=1}^{n}\sum_{y=1}^{m} A_i(x)B_j(y)} \tag{2}$$

Here the image $I(x,y)$ is compressed to $F_{i,j}$ by using a discrete F-transform in two variable as $F_{i,j}$. Then inverse fuzzy F-transform can also be defined as:

$$R_{n,m}(i,j) = \sum_{k=1}^{n}\sum_{l=1}^{m} F_{kl}A_k(i)B_l(j) \tag{3}$$

2.2 Adaptive Quantization Coding (*AQC*) and Intensity Based AQC (*IBAQC*) Algorithm

Adaptive Quantization Coding (AQC) as it name signifies, is such that quantization levels are adaptively changed to compute bit blocks. AQC compression algorithm was initially introduced for reduction of motion blur in the liquid crystal display (LCD). The AQC algorithm coding steps are:

1. Divide the image into non-overlapped blocks of $(m \times n)$ pixels
2. Compute the quantizer step using the following relation:

$$Qt_{step} = \frac{(Bl_{max} - Bl_{min})}{Q_L} \tag{4}$$

   Where Qt_{step} is the quantizer step, and Qt_L represents the number of quantization levels, and Bl_{max} and Bl_{min} are the maximum and minimum of the each block, respectively
3. Compute the bit plane (Bit_{plane}) of each block using the quantizer step according to the following relation:

$$Bit_{plane} = round\left[\frac{(Im_{in} - Bl_{min})}{Qt_{step}}\right] \tag{5}$$

4. The compressed data finally includes a bit plane, quantizer step and the minimum of each block.

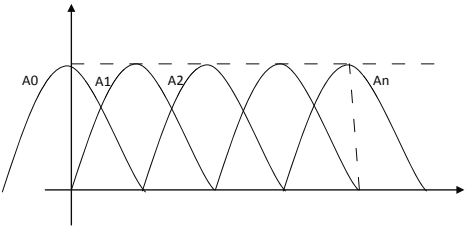


Fig. 1. Fuzzy partition functions

The AQC algorithm decoding steps are:

1. The decoding process is to compute the value of decoded image for each block as:

$$Im_{decode} = Bl_{min} + Bl_{plane} \times Qt_{step} \tag{6}$$

According to the algorithm described, the bits per pixels rate of a grey scale image compressed using this algorithm can be computed as:

$$bpp = \frac{((bits\ for\ step) + (bits\ for\ plane) + (bits\ for\ min))}{m \times n} \tag{7}$$

The typical value of a grey scale lena image with (512×512) pixels with a block of (4×4) pixels and 8 quantization levels, results in bpp as 3.82.

An advanced version of this AQC algorithm presented in [7] as Intensity based Adaptive Quantization Coding IBAQC with the aim to check the intensity variations in each block. Here the image was first divided into non-overlapped blocks of size $m \times n$. And

for each block the maximum, minimum in the block and quantizer step was computed according to AQC algorithm. Here an already set threshold is used to classify the block as either low variation smooth block or the high variation edge block on the basis of value of Qt_{step} calculated in AQC. This is because the smooth blocks required lower quantization steps compared with the edge blocks. In IBAQC there is no need to send the complete Bit plane in the case of low variation only the one single value i.e. Bl_{min} is sufficient to reconstruct this block. Thus the bit plane and quantization step bits of block where the quantization step Qt_{step} was less than the pre-defined threshold were discarded and only the minimum value was encoded, while for all other blocks the complete information was added to coded data. This process results to lower bit rate as compared to AQC but obviously on the trade off with some reconstruction quality. The decoder process of IBAQC is similar to AQC but contrasts as this is sensitive to the flag bit. If it is in high state the block is classified edge block as decoded as AQC but if it is low it is decoded as smooth block and its all values will be set to Blmin. The complete Block diagram of IBAQC is:

3 Proposed IBAFC Algorithm

The proposed IBAFC algorithm depends upon fuzzy F-Transform and Intensity based Adaptive Quantization Coding IBAQC. In IBAQC a drawback is that the bit plane is computed as some value of quantization threshold step (Qt_{step}) i.e. most of the correlation information between pixels is lost at this step because of quantization value. Hence to obtain the optimum result, the bit plane of the image is extracted with fuzzy set logic i.e. the fuzzy bit plane with correlation between the pixels maintained at fewer bit representation.

The proposed IBAFC algorithm is:

1. Divide the image into non-overlapped blocks of $(m \times n)$ pixels
2. Compute the quantizer step using the following relation:

$$Qt_{step} = \frac{(Bl_{max} - Bl_{min})}{Q_L} \tag{8}$$

   Where Qt_{step} is the quantizer step, and Qt_L represents the number of quantization levels, and Bl_{max} and Bl_{min} are the maximum and minimum of the each block, respectively

   Thus to decide about a block whether it is low variation smooth or high variation block edge block, we use

   (a) If $Qt_{step} < threshold$ the block is classified as smooth block i.e. low intensity variation block.

   (b) If $Qt_{step} > threshold$ the block is classified as edge block i.e. high intensity variation block.

3. For the low variation block, Compute mean of block as Bl_{mean} and set Flag = 0. And for the high variation block, the fuzzy bit plane is extracted of the block according to the fuzzy F-Transformation (FTR) defined in [6] and set $Flag = 1$.

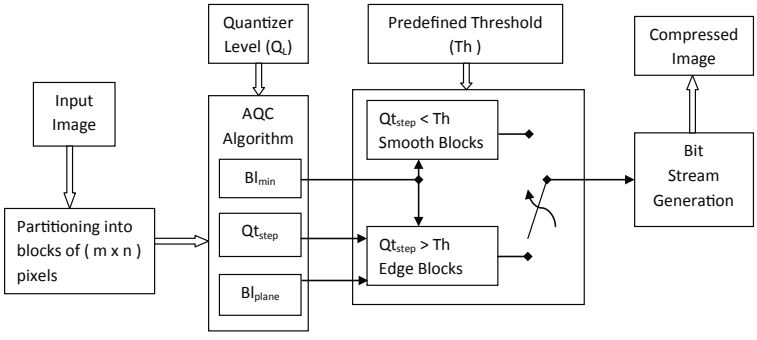


Fig. 2. IBAQC Encoder [7]

4. For each flag bit
   - if it is false
     code the data as Bl_{mean} i.e the mean value of the block
   - and if it is true
     code the data as reduced fuzzy bit plane.

The block diagram of proposed IBAFC coding process is described in fig 3.

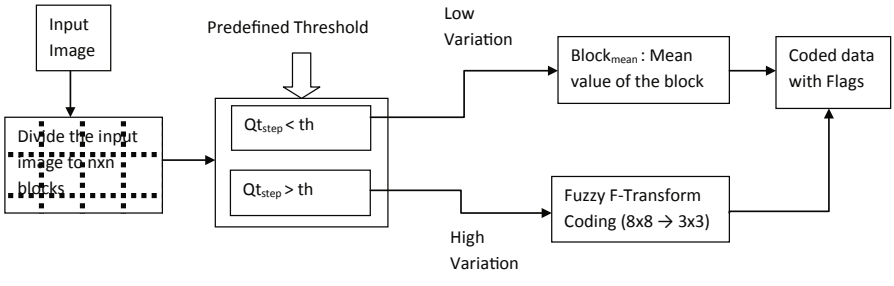


Fig. 3. Proposed IBAFC Algorithm

4 Simulation Results and Discussion

It is quite clear that the combination of Adaptive Quantization with fuzzy F-transform based image compression will results in lossy data compression method, when compared to original image. In this process the loss is due to both the mean value computation as well as due to fuzzy F-Transform based coding. Thus if this method has to be near optimal image compression system the prime requirement is of perfect classification of blocks. Compared to IBAQC [7] in this proposed method of IBAFC, the inclusion of fuzzy F-Transform results in fuzzy bit block compared to AQC block. The

Table 1. MSE, PSNR and bpp Comparisonfor blocks size 2×2 Predefined Threshold set at 10

	AQC			IBAQC			IBAFC		
Block Size 2×2	PSNR (dB)	MSE	bpp	PSNR (dB)	MSE	bpp	PSNR (dB)	MSE	bpp
Lena	-	-	-	30.62	56.22	4.78	31.19	49.44	4.27
Goldhill	-	-	-	32.09	40.11	4.98	33.87	26.67	4.44
Peppers	-	-	-	32.41	32.25	4.84	34.49	23.13	4.13

Table 2. MSE, PSNR and bpp Comparisonfor blocks size 4×4 Predefined Threshold set at 10

	AQC			IBAQC			IBAFC		
Block Size 4×4	PSNR (dB)	MSE	bpp	PSNR (dB)	MSE	bpp	PSNR (dB)	MSE	bpp
Lena	43.71	2.77	4.92	25.76	172.62	3.82	26.27	153.49	3.76
Goldhill	47.28	1.21	5.15	23.85	267.46	3.94	25.11	200.48	3.39
Peppers	33.91	26.48	4.36	24.34	238.97	3.61	26.01	162.95	3.57

Table 3. MSE, PSNR and bpp Comparisonfor blocks size 8×8 Predefined Threshold set at 10

	AQC			IBAQC			IBAFC		
Block Size 8×8	PSNR (dB)	MSE	bpp	PSNR (dB)	MSE	bpp	PSNR (dB)	MSE	bpp
Lena	43.51	2.87	4.28	23.22	310.59	3.02	23.99	259.47	1.56
Goldhill	42.24	3.87	5.35	20.70	554.31	3.15	22.84	338.12	1.23
Peppers	41.78	4.31	4.17	20.77	543.75	3.11	21.35	476.52	1.49

advantage results due to this is that the correlation between the pixels is also maintained while having very good compression. The experimental results of the proposed method are presented subjectively and objectively. Experiments were conducted using the images lena, peppers and goldhill of size 512×512 with 256 gray levels. The objective metric used for comparison of original image I and decoded image $\hat{I}$ is the PSNR (peak signal-to-noise ratio) and the MSE (mean square error) value. In general, the larger PSNR(dB) value, the better is the reconstructed image quality. The mathematical formulae for the computation of MSE & PSNR is

$$MSE = \sum_{m=1}^{M} \sum_{n=1}^{N} \left[I(m,n) - \hat{I}(m,n) \right]^2 \tag{9}$$

$$PSNR = 10 \log_{10} \left(\frac{255^2}{MSE} \right) \tag{10}$$

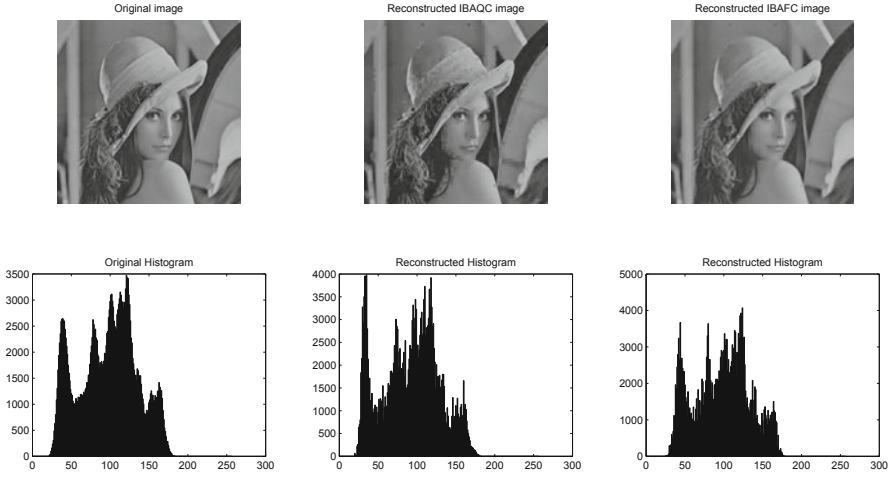


Fig. 4. Original Lena and reconstructed with IBAQC and proposed IBAFC

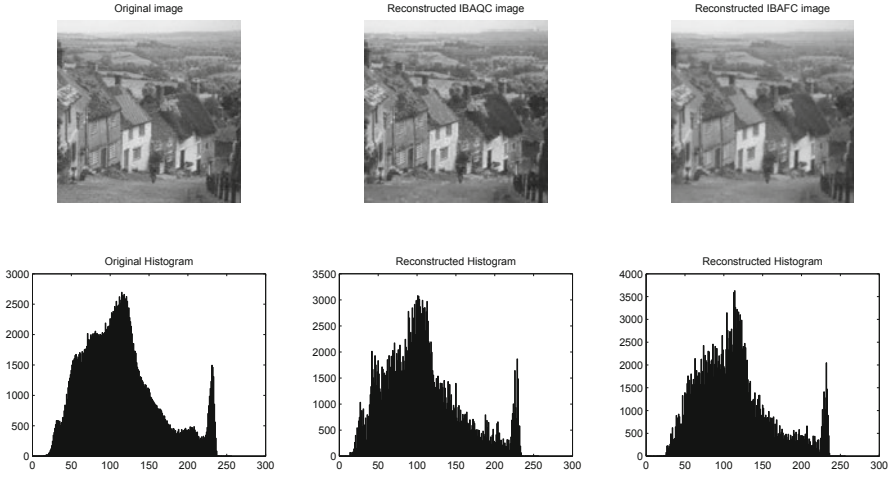


Fig. 5. Original Goldhill and reconstructed with IBAQC and proposed IBAFC

From the results it is very clear that, usually it is not worth to have AQC for very small block sizes say i.e. 2×2. Because AQC for small sizes provides results approximatley same as image data but IBAFC proves good as compare to AQC and IBAQC. For block sizes of 4×4 and above the IBAFC provides good results superior to IBAQC with more PSNR at reduced bpp.

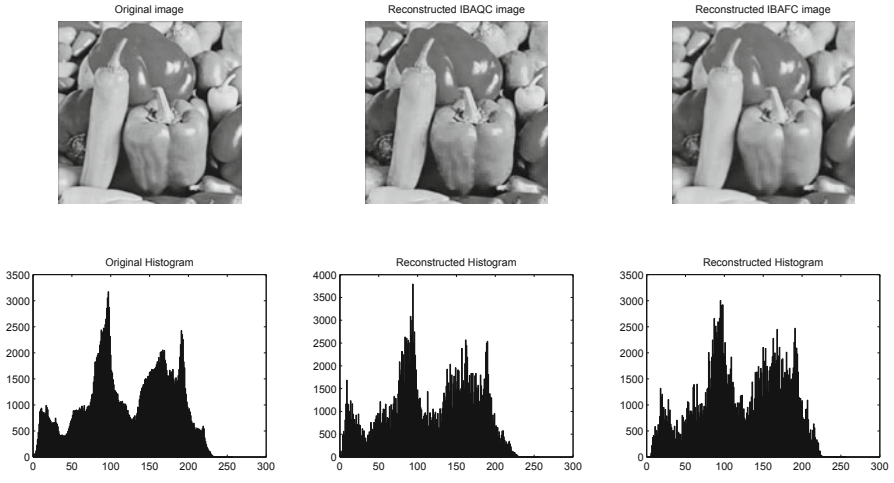


Fig. 6. Original Peppers and reconstructed with IBAQC and proposed IBAFC

5 Conclusion

The IBAFC algorithm has been introduced in this paper. A kind of fuzzy based bit block is obtained in proposed IBAFC algorithm as adaptive quantized bit block in IBAQC. The results proves that this fuzzy bit block based image compression algorithm gets effective performance in terms of compression rate and better reconstruction PSNR quality. It is also being proved that the IBAFC results in fuzzy bit block that not only provides more compression as well as it maintains corellation between the pixels too and above all, the PSNR obtained by IBAFC was better than that conventional AQC and IBAQC algorithms for less bpp.

References

1. El-Khamy, et al.: A fuzzy gradient-adaptive lossy predictive coding technique. In: Proceedings of the IEEE Twentieth National Radio Science Conference, NRSC, pp. C–5 1–C–5 11 (2003)
2. Martino, F.D., et al.: Image coding/decoding method based on direct and inverse fuzzy transforms. Elsevier International Journal of Approximate Reasoning 48, 110–131 (2008)
3. Jeung, Y.C., Wang, J., Min, K.Y., Chong, J.W.: Improved btc using luminance bitmap for color image compression. In: Proceedings of the 2009 2nd International Congress on Image and Signal Processing, CISP 2009 (2009)
4. Guo, J.M.: Improved block truncation coding using modified error diffusion. Electronics Letters 44, 462–464 (2008)
5. Rayon, P., Osslan, V., Villegas, O.V., Elias, R.P., Salazar, A.M.: Edge preserving lossy image compression with wavelets and contourlets. In: Proceedings of the IEEE Int. Conference on Electronics, Robotics and Automotive Mechanics, CERMA, vol. 01, pp. 3–8 (2006)
6. Perfilieva, I.: Fuzzy transforms. Fuzzy Sets and Systems 157(8), 993–1023 (2006)

7. Yakovlev, A., Al-Azawi, S., Boussakta, S.: Performance improvement algorithms for colour image compression using dwt and multilevel block truncation coding. In: Proceeding of IEEE Intl. Conf. CSNDSP (2010)
8. Abraham, T.M., Amarunnishad, T.M., Govindan, V.K.: Improving btc image compression using a fuzzy complement edge operator. Elsevier Signal Processing Journal 88(1), 2989–2997 (2008)
9. Abraham, T.M., Amarunnishad, T.M., Govindan, V.K.: A fuzzy complement edge operator. In: IEEE Proceedings of the Fourteenth International Conference on Advanced Computing and Communications, Mangalore, Karnataka, India (December 2006)
10. Masaki, I., Horn, B.K.P., Desai, U.Y., Mizuki, M.M.: Edge and mean based compression. MIT Artificial Intelligence Laboratory AI (Memo No.1584), pp. 533–536 (November 1996)

Periocular Feature Extraction Based on LBP and DLDA

Akanksha Joshi, Abhishek Gangwar, Renu Sharma, and Zia Saquib

Center for Development of Advanced Computing, Mumbai, India
{akanksha,abhishekg,renu,saquib}@cdac.in

Abstract. Periocular recognition is an emerging field of research and people have experimented with some feature extraction techniques to extract robust and unique features from the periocular region. In this paper, we propose a novel feature extraction approach to use periocular region as a biometric trait. In this approach we first applied Local Binary Patterns (LBPs) to extract the texture information from the periocular region of the image and then applied Direct Linear Discriminant Analysis (DLDA) to produce discriminative low-dimensional feature vectors. The approach is evaluated on the UBIRIS v2 database and we achieved 94% accuracy which is a significant improvement in the performance of periocular recognition.

Keywords: Periocular recognition, local binary patterns, direct linear discriminant analysis.

1 Introduction

Over the past few years, iris recognition has gained a lot of prominence. Other ocular traits like retina, sclera and conjunctival vasculature have also been identified. In spite of the tremendous progress, the main challenge in ocular biometrics is the recognition in non-constrained environment. Recently a new area referred as periocular is gaining lot of attention of researchers.

The periocular region is the part of the face immediately surrounding the eye [1] (may contain eyebrows also). Its features can be classified into two categories global and local. Global features include the upper/lower eye folds, the upper/lower eyelid, wrinkles, and moles. Local features are more detailed and include skin textures, pores, hair follicles, and other fine dermatological features. Some authors have published papers showing that periocular region can be used as soft biometrics [12], as well as a secondary method supporting the primary biometric like iris recognition when the primary biometric is not available [5, 6]. Recently few authors have also published their work to support its use as primary biometric modality [1, 3, 4]. Park et al. in [1] used LBPs (Local Binary Patterns), GO (Gradient Orientation) histograms, SIFT (Scale Invariant Feature Transform) on a dataset of 899 visible-wavelength image and reported accuracies 70% - 80%. Miller et al. in [4] used LBPs on the images from the FRGC and FERET databases. They reported 89.76% accuracy on FRGC and 74.07% on the FERET. Adams et al. in [6] used LBPs and GEFE (Genetic & Evolutionary Feature Extraction) for optimal feature selection and reported increase in accuracy from 89.76% to 92.16% on FRGC dataset and 74.04% to 85.06% on the FERET dataset.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 1023–1033.
springerlink.com

In this paper, we propose a new feature extraction technique using LBPs combined with DLDA. DLDA extracts discriminative low-dimensional feature vectors from periocular region while utilizing all the advantages of LBPs. We evaluated our approach on a larger dataset of images from UBIRIS v2 [2] database which contains the images captured in more realistic situations like, in the visible wavelength, at-a-distance (between four and eight meters) and on on-the-move. UBIRIS v2 is originally created for iris recognition systems but images also contain periocular region with varying area, scale, illumination, and noises (Figure 3). Therefore we utilized it to evaluate our approach and we achieved 94% accuracy, which is a significant improvement in the performance of periocular recognition.

The remainder of this paper is as follows. Section 2 provides a brief overview of periocular recognition. Section 3 explains feature extraction, LBPs and DLDA. Section 4 explains our proposed approach. In section 5, we described how image dataset was created and in section 6 we have explained the experiments and results. In Section 6, conclusion and future work is given.

2 Periocular Recognition

Periocular Recognition starts with capturing the iris along with the surrounding region or face and then segmentation of the region of interest (periocular region) from the image. In [4, 5, 6] authors have also proposed to mask iris and surrounding sclera part to prevent the iris and sclera texture. Then global or local feature extraction techniques are applied to extract periocular features. Global features are extracted from a common region of interest from the given image hence image is aligned by using iris, eyelids or eye position. Local features are extracted by detecting local key points just like in fingerprint recognition systems. Once features are extracted, then a suitable matching criteria is applied to find similarity in the periocular images.

3 Periocular Feature Extraction

For periocular feature extraction, authors have proposed different techniques like Local binary Patterns [1, 4, 6] and Gradient Orientation histograms [1] to extract global features and SIFT [1] to find the local key points feature from the image. In our approach we have extracted global information from periocular region.

3.1 Local Binary Patterns

LBPs, first introduced by Ojala et al. [14] quantify intensity patterns and are helpful in detecting patterns like spots, line edges, corners and other texture patterns from the image. LBPs have been successfully applied to face recognition [9] facial expression recognition [11], gender classification [13], iris [15] and palm print recognition [16]. The features extracted using LBP can be encoded into a histogram in the form of bins. The detailed LBPs description can be found in [14]. A brief description of LBPs is given below.

In original LBP, 3*3 neighborhood is thresholded using the value of center pixel as shown in Figure.1. The pixels having values greater than the center pixel is encoded as 1 and pixels having values less than the center pixel are encoded as 0 and final LBP

string is generated by concatenating them. Thus, LBP operator to be applied over image is defined as

$$LBP_{N,R}(p,q) = \sum_{i=0}^{N-1} s(n_i - n_c)2^i \tag{1}$$

$$s(.) = \begin{cases} 1 & if \quad n_i - n_c \geq 0 \\ 0 & otherwise \end{cases} \tag{2}$$

Where n_c is gray level intensity of center pixel and n_i is gray level intensity of neighboring pixels. A uniformity measure for LBPs is defined which corresponds to number of 0/1 of bitwise or sign changes in the pattern. A uniform LBP operator considers only those LBP strings in which there are at most two bitwise transitions from 0 to 1 or vice versa. For example: 11110111, 11111110 are uniform patterns, whereas 11010111, 10110101 are non uniform patterns. There are 58 possible levels of uniform patterns and the rest 198 labels contribute for non uniform patterns, whose values can be stored in the 59th label. Thus, with uniform LBP operator, we can encode each block in our image using 59 bins.

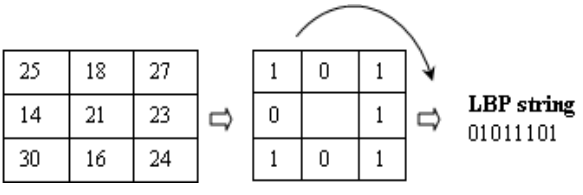


Fig. 1. Obtaining LBP code for a 3*3 pixel neighborhood

Ojala et al. [14] also proposed rotation invariant local binary patterns based on a circularly symmetric neighbor set of P members on a circle of radius R, denoting the operator as $LBP_{P,R}^{riu2}$. The rotation invariant LBP operator is an excellent measure of the spatial structure of local image texture, but by definition, it discards the other important property of local image texture, i.e., contrast, since it depends on the gray scale.

3.2 Direct Linear Discriminant Analysis (DLDA)

Direct Linear Discriminant Analysis [7] is a well known classification technique and has been applied over face to reduce the feature dimensions. It basically transforms the data to lower dimensions, without losing the discrimination information. Previously, people applied PCA + LDA approach [10] but applying PCA tend to lose the discriminatory information among classes and thus yielding low accuracy. LDA aims at maximizing the ratio of between class scatter S_b and within class scatter S_w. It discards the null space of S_w, which contains the most discriminatory information according to Chen et al. [8]. The key idea of DLDA algorithm is to discard the null space of S_b,

which contains no useful information rather than discarding the null space of S_w, which contains the most discriminative information. This can be done by first diagonalizing S_b and then diagonalizing S_w. The whole DLDA algorithm is outlined below.

1. First diagonalize the S_b matrix, such that

$$V^T S_b V = D \tag{3}$$

It involves finding the eigenvectors of matrix S_b and matrix D contains the corresponding eigen values. We discard those values from V which contains eigen values corresponding to 0 such that

$$Y^T S_b Y = D_b > 0 \tag{4}$$

Where Y is n*m matrix (n is the feature dimension) and contains first m columns from V and D_b is m*m matrix corresponding to non-zero eigen values.

2. Let $Z = YD_b^{-1/2}$, where

$$Z^T S_b Z = I \tag{5}$$

Where, Z unitizes S_b and reduces dimensionality from n to m.

3. Now we need to diagonalize $Z^T S_w Z$ as,

$$U^T \left(Z^T S_w Z\right) U = D_w \tag{6}$$

4. Let $A = U^T Z^T$ diagonalizes both numerator and denominator in Fisher's criteria as, $A^T S_w A = D_w$, $A^T S_b A = I$

5. Thus, we get the final transformations as

$$T = D_b^{-1/2} AX \tag{7}$$

Where X is the feature vector extracted from the image and T is the reduced feature vector.

4 Proposed Approach

The LBP operator is good at texture classification, and to extract texture information from the images, we have chosen uniform LBP operator. The LBP feature extraction process is shown in Fig. 2. We divided the entire periocular image into equal sized

blocks and then LBP is applied at each block. The histogram features extracted from each block is concatenated to form a single feature. Then DLDA is applied to obtain discriminative low dimensional feature representation. Euclidean distance is used for finding similarity among the feature vectors. If the feature vectors for k classes are -

$$X^m = \left[x_1^m, x_2^m \ldots x_n^m\right] \qquad \text{Where, } 1 \le m \le k \text{ ,}$$

Then the test feature vector X is assigned to the $class_i \in \{class_1, class_2 ..., class_k\}$ with minimum euclidean distance as

$$\text{Class}_i = \min_m (ED(X, X^m)) \tag{8}$$

Where ED is the euclidean distance between X and X^m.

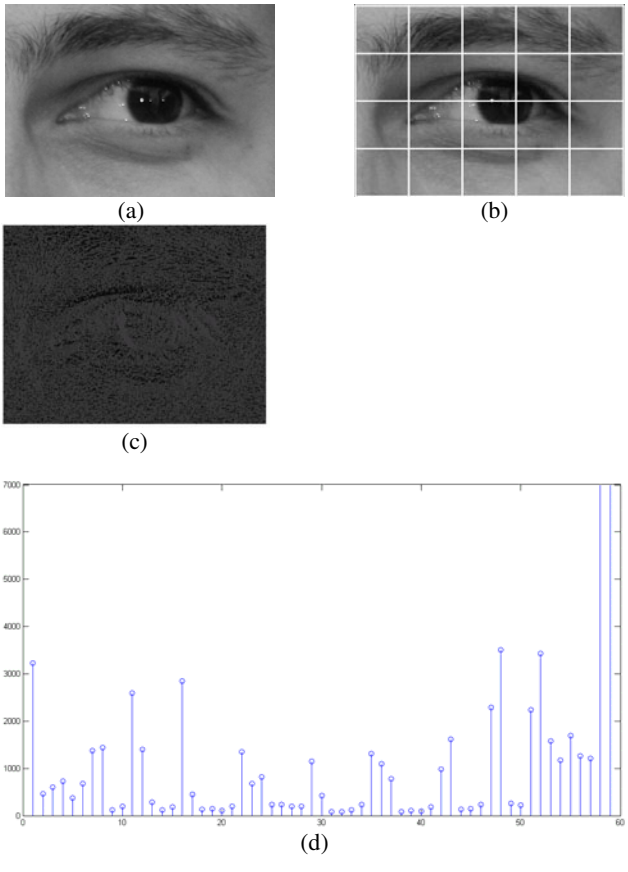


Fig. 2. (a) Eye image converted to gray level from UBIRIS v2 database (b) Image divided into blocks (c) LBP scores (d) Histogram for a block of image

5 Database

There is no public database available for periocular region images. We found that recently released UBIRIS v2 [2] database contains images of iris along with the surrounding region (which can contribute for periocular recognition). The original intent of the database is the development of robust iris recognition algorithms in visible spectrum. The images are captured during multiple sessions in non-constrained conditions (at-a-distance, on-the-move and in the visible wavelength). UBIRIS v2 database contains total 11102 images of 261 subjects (we found that images of 2 subjects were not given in download). The right eye images of a user captured in one session from the UBIRIS database is shown in Figure 3.

In UBIRIS v2 for each subject 30 or 60 images are taken on an average at distance of 4 to 8 meters and thus contain a lot of variation. Some images do not contain the region surrounding eye (periocular information) and are much focused towards the iris; therefore we have not included such images for our experiments. To test left and right eye separately for periocular recognition we created different database for left and right eyes containing periocular region. We selected images for our experiments as given in Table 1.

Table 1. Image Dataset for Periocular Recognition

Session	No. of subjects	Actual images In UBIRIS v2		Images selected for periocular recognition experiments		
		LE	RE	LE	RE	Total
S1	259	15	15	6	6	3108
S2	111	15	15	6	6	1332
						4440

Note: LE = Left Eye, RE = Right Eye
S1=Session one, S2=Session Two
Total= [Selected (LE + RE) * No. of Subjects]

6 Experimental Evaluation

For evaluating our approach, we selected 6 left and 6 right eye images for each user from two sessions; total 4440 images from UBIRIS database as explained in section 5. Each image is given an id e.g. P_1L_1 to P_1L_6 for first users' left eyes and P_2R_1 to P_2R6 for second users' right eyes. Evaluation is conducted on 12 sets of images. Set_1 to set_6 are for left eye dataset and set_7 to set_{12} are for right eye dataset. The images in different sets are selected as follows,

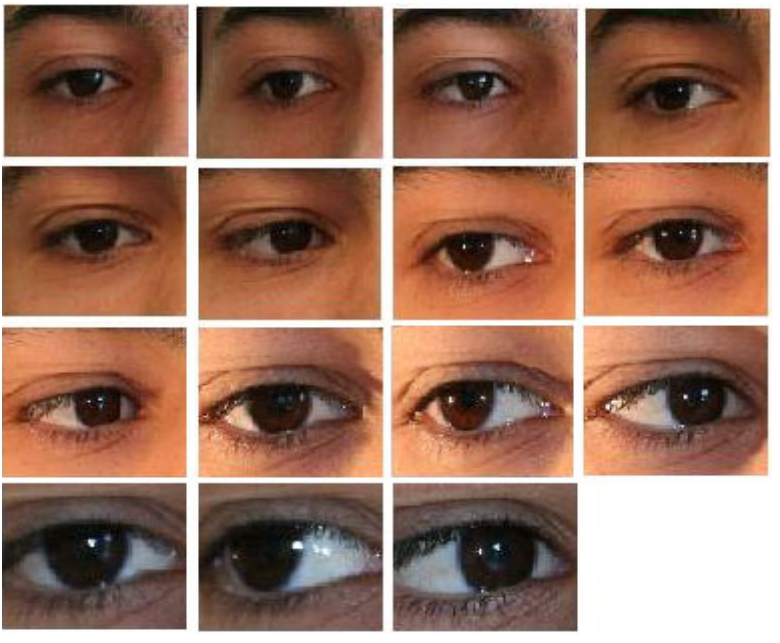


Fig. 3. Images of a person from the UBIRIS v2 database

$$\text{Set}_i = \begin{cases} \text{probe} = \{P_n L_M\},\ \text{gallery} = \{P_n L_{(S-M)}\}\ if\ i \le 6 \\ \text{probe} = \{P_n R_M\},\ \text{gallery} = \{PnR_{(S-M)}\}\ if\ i > 6 \end{cases} \tag{9}$$

Where,

i is set id and $1 \le i \le 12$,

n is person id and $1 \le n \le 259$,

S is total images for a user and S = {1, 2, 3, 4, 5, 6}

and $M = \begin{cases} i & if\ i \le 6 \\ i-6 & if\ i > 6 \end{cases}$

For example $\mathbf{Set_3}$ will contain

Probe= $\{P_1L_3, P_2L_3 \ldots\ldots P_{259}L_3\}$ and

Gallery= $\{P_1L_1, P_1L_2, P_1L_4, P_1L_5, P_1L_6 \ldots\ldots$
$\ldots..P_{259}L_1, P_{259}L_2, P_{259}L_4, P_{259}L_5, P_{259}L_6\}$

Table 2. Rank-1 Identification Accuracies using LBP and LBP + DLDA for Different Sets of Left Eye Images

Training set (Left eye)	Rank-1 Identification accuracy (%) using LBP		Rank-1 Identification accuracy (%) using LBP + DLDA	
	(S1)	(S2)	(S1)	(S2)
Set_1	**79.92**	63.06	**95.37**	90.99
Set_2	76.00	66.66	90.73	90.99
Set_3	62.54	67.56	84.17	94.49
Set_4	76.40	**79.27**	92.60	**96.39**
Set_5	74.13	71.17	92.28	96.39
Set_6	72.20	72.07	90.73	94.59

Table 3. Rank-1 Identification Accuracies using LBP and LBP + DLDA for Different Sets of Right Eye Images

Training set (Right eye)	Rank-1 Identification accuracy (%) using LBP		Rank-1 Identification accuracy (%) using LBP + DLDA	
	(S1)	(S2)	(S1)	(S2)
Set_7	77.99	83.78	93.03	94.59
Set_8	68.72	73.87	86.87	93.69
Set_9	74.10	75.67	89.18	89.18
Set_{10}	**79.10**	**83.78**	**93.82**	**97.29**
Set_{11}	69.88	77.47	89.86	89.18
Set_{12}	69.40	81.08	88.41	91.89

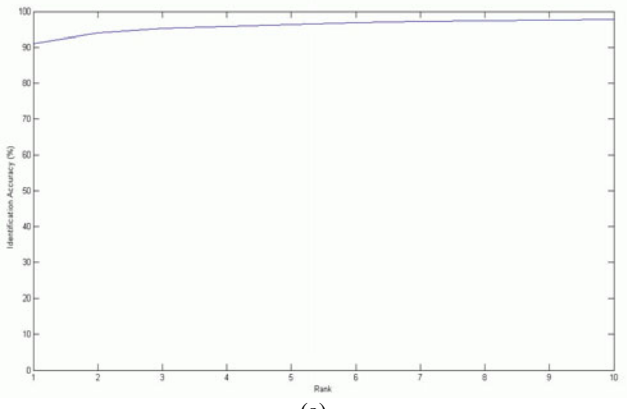
We performed the following experiments on this dataset.

Experiment 1: First we applied LBPs for each of the gallery probe pairs and used CityBlock distance [4] as matching criteria for finding similarity among the LBPs feature vectors. We achieved accuracies upto 79.92% and 79.10% for left and right eye datasets for session one, whereas 79.27% and 83.78% for left and right eye dataset for session two. The rank-1 identification accuracy using LBPs for left and right eye data on each of the 12 sets for both sessions are given in Table 2 and 3.

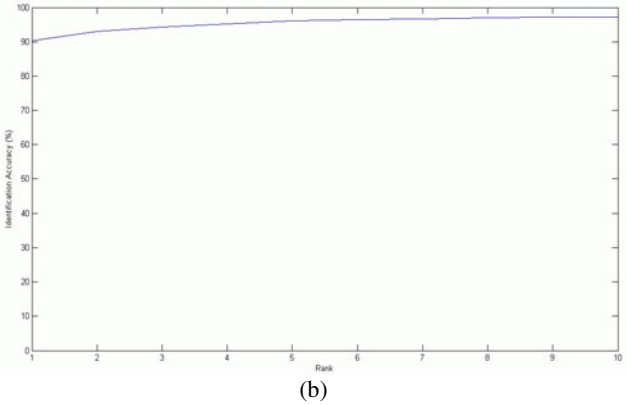
Experiment 2: Then we evaluated our approach (LBPs + DLDA) on each of the twelve gallery probe pairs. We achieved upto 95.37% rank-1 identification accuracy for left eye and 93.82 % for right eye UBIRIS v2 session one database and for

session two we got accuracies upto 96.39% and 97.29 for left and right eye datasets. As seen from Table 2 and 3 LBPs along with DLDA performs much better as compared to LBPs alone. The CMC curves for left and right eye periocular recognition for our approach are shown in Fig. 4.

Experiment 3: We also experimented with rotation invariant LBPs on our dataset. We achieved average accuracy of 66.34% accuracy for left eye dataset and 64.66% for right eye dataset. For rotational invariant LBPs combined with DLDA, we achieved average accuracy of 79.02 % accuracy for left eye dataset and 75.41% for right eye dataset.



(a)



(b)

Fig. 4. CMC curve for periocular recognition using left (a) and right eye (b)

Experiment 4: In this experiment we created gallery from session one images and probe from session two images. We achieved an average accuracy of 92.14 for left eye and 93.89 for right dataset.

7 Conclusion and Future Work

In this paper, we have presented a novel feature extraction technique for periocular recognition. The approach is evaluated on the images captured in realistic and non-constrained environment (visible spectrum, on-the-move and at-a-distance). We first extracted features using Local Binary Patterns and then applied DLDA to produce discriminative low dimensional feature vectors. Our experimental results show that combining LBPs and DLDA gives better performance than LBPs alone and feasibility of using periocular region as biometric modality. The future work will involve the study of make-up effects and age related changes in the periocular region and identifying more robust feature extraction techniques.

References

[1] Park, U., Ross, A., Jain, A.K.: Periocular biometrics in the visible spectrum: A feasibility study. In: Proc. IEEE Int. Conf. on Biometrics: Theory, Applications, and Systems (BTAS 2009), pp. 1–6 (September 2009)

[2] Proença, H., Filipe, S., Santos, R., Oliveira, J., Alexandre, L.A.: The UBIRIS. v2: A database of visible wavelength images captured on-the move and at-a-distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 99(Rapid Posts) (2009)

[3] Bharadwaj, S., Bhatt, H.S., Vatsa, M., Singh, R.: Periocular biometrics: When iris recognition fails. In: 2010 Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS), September 27-29, pp. 1–6 (2010)

[4] Miller, P., Rawls, A., Pundlik, S., Woodard, D.: Personal identification using periocular skin texture. In: Proc. ACM 25th Symposium on Applied Computing (SAC 2010), pp. 1496–1500 (2010)

[5] Woodard, D.L., Pundlik, S., Miller, P., Jillela, R., Ross, A.: On the fusion of periocular and iris biometrics in non-ideal imagery. In: Proc. Int. Conf. on Pattern Recognition (2010)

[6] Adams, J., Woodard, D.L., Dozier, G., Miller, P., Bryant, K., Glenn, G.: Genetic-based type II feature extraction for periocular biometric recognition: Less is more. In: Proc. Int. Conf. on Pattern Recognition (2010)

[7] Yu, H., Yang, J.: A Direct LDA Algorithm for High-Dimensional Data with Application to Face Recognition Interactive System Labs. Carnegie Mellon University, Pittsburgh

[8] Chen, L., Liao, H., Ko, M., Lin, J., Yu, G.: A new lda-based face recognition system which can solve the small sample size problem. Pattern Recognition 33(10), 1713–1726 (2000)

[9] Heusch, G., Rodriguez, Y., Marcel, S.: Local binary patterns as an image preprocessing for face authentication. In: Proc. of International Conference on Automatic Face and Gesture Recognition, pp. 9–14 (2006)

[10] Swets, D., Weng, J.: Using discriminant eigenfeatures for image retrieval. PAMI 18(8), 831–836 (1996)

[11] Liao, S., Fan, W., Chung, A., Yeung, D.: Facial expression recognition using advanced local binary patterns, tsallis entropies and global appearance features. In: Proc. of the IEEE International Conference on Image Processing (ICIP), pp. 665–668 (2006)

[12] Lyle, J.R., Miller, P.E., Pundlik, S.J., Woodard, D.L.: Soft biometric classification using periocular region features. In: 2010 Fourth IEEE International Conference on Biometrics: Theory Applications and Systems (BTAS), September 27-29, pp. 1–7 (2010)
[13] Sun, N., Zheng, W., Sun, C., Zou, C., Zhao, L.: Gender Classification Based on Boosting Local Binary Pattern. In: Wang, J., Yi, Z., Żurada, J.M., Lu, B.-L., Yin, H. (eds.) ISNN 2006. LNCS, vol. 3972, pp. 194–201. Springer, Heidelberg (2006)
[14] Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 24(7), 971–987 (2002)
[15] Sun, Z., Tan, T., Qiu, X.: Graph Matching Iris Image Blocks with Local Binary Pattern. In: Zhang, D., Jain, A.K. (eds.) ICB 2005. LNCS, vol. 3832, pp. 366–372. Springer, Heidelberg (2005)
[16] Wang, X., Gong, H., Zhang, H., Li, B., Zhuang, Z.: Palmprint identification using boosting local binary pattern. In: Proc. 18th International Conference on Pattern Recognition, ICPR (2006)

Towards XML Interoperability

Sugam Sharma[1], S.B. Goyal[2], Ritu Shandliya[3], and Durgesh Samadhiya[4,*]

[1] Dept. of Comp. Science, Iowa State University, USA
[2] Dept. of Comp. Science & Engg., MRCE, Faridabad, India
[3] Dept. of Comp. Science & Engg., Shobhit University, India
[4] Department of Information Technology, Chung Hua University, Taiwan
goyalsb@yahoo.com,
samadhiya.durgesh@gmail.com

Abstract. Now a day's distributed computing has become a common phenomenon. As the system designed from the bottom up with networking in mind, distributed computing makes it very easy for computers to cooperate and today, scientific world is nourishing the benefits provided by distributing computing under the broad umbrella of client server architecture. Information is requested and used at various distant physical or logical locations as per the requirement by the users. This is very unlikely that all the engaged users use the same computing environment. XML (Extensive Markup Language) technology has emerged as an efficient medium for information transfer and attempting to offer similar kind of information environment up to an extent by using its potential property called, interoperability. Interoperability is a kind of ability of software and hardware on different machines from different vendors to share data. In this paper we employ a few criteria to evaluate interoperability of XML in heterogeneous distributed computing environment.

Keywords: XML, heterogeneous, distributed, environment, interoperability, evaluation criteria.

1 Introduction

The interoperability is the ability of two or more systems or components to exchange information and to use the information that has been exchanged. Interoperability is a property referring to the ability of diverse systems and organizations to work together. It is a key challenge in the realms of the Internet of Things [15]. XML (Extensive markup language) is software and hardware independent technology is considered as one of the standard medium for transmitting data as XML has become one of primary standards for data exchange over Web [16]. Once a web service is produced by a producer, it can be consumed anywhere, any number of times, in heterogeneous distributed environment and the scientific world is harnessing these benefits of web services at various levels. For example a naive user – non technical- can get the benefits of web services in web sites (www) at remote nodes, consumption of web service by software developer in developing new application, or consumption by scientists to explore further research.

* Corresponding author.

D.C. Wyld et al. (Eds.): Advances in Computer Science, Eng. & Appl., AISC 166, pp. 1035–1043.
springerlink.com

In a distribute environment it is very unlikely that all the engaged users use the same computing environment. In other words, the presence of the heterogeneous distributed environment is more likely. The concept of web services introduces any new concept of reusability. Which means, the produced web services can be reused in various application which leads to the concept of Software Product Line Engineering (SPLE) [13]? As per the SPLE the reusability always helps in cost reduction and reduction in Time to market (TTM) [14]. The further description of SPLE is beyond the scope of this paper. Thus the concept of web services helps in coordination among heterogeneous distributed computing nodes to work together. Due to the interoperable potential of XML, caters the nature of heterogeneous distributed computing environment. Interoperability helps the heterogeneous systems to work with each other in a close coordination bypassing any intensive implementation or restricted access. XML interoperability tailors the exposed interfaces of the system to be completely understood at the receiving heterogeneous computing nodes. In literature system interoperability has been defined in various flavors. IEEE glossary [6] explains interoperability as the ability of two or more systems that exchange information and use it. In this paper we devise few hypotheses of evaluation criteria to evaluate XML interoperability: 1) Support for Object Relational Data, 2) Intactness of Existing Systems, 3) Heterogeneity of Spatial Validity,4) Information Exchange via Middle Tier Format Changers. Later in this paper, we have evaluated whether XML interoperability supports the hypothesis.

The rest of the paper is organized as follows. Section 2 describes background research work. Section 3 elaborates about XML. The system interoperability has been described in section 4 in sufficient detail. In section 5 we explain about evaluation criteria. Section 6 throws intense light on XML interoperability evaluation criteria and the paper is concluded in section 7.

2 Extensive Markup Language

XML is widely used for promoting interoperability and data integration and transportation, it has some interoperability challenges that need to be addressed. XML is a semi-structured data representation and today widely used in heterogeneous distributed environment for information exchange in much flexible tag based format. It was initially designed to address the issues in publishing voluminous data and mainly used as the customization, but lately its potential has been infused for exchange and transportation of a huge variety of information among heterogeneous distributed computing environment. And because of this property of XML, the web world is intensively using it using an artifact, called web service. In the web world which is considered as the distributed environment, the designing of the web services is built on XML as the underneath building block (Figure 1). Extensive markup language (XML) is a common W3C standard for semi-structured data representation, being used by web and other applications for data exchange. It is considered as one of the simplest and flexible text format being used widely today, which makes it easier to work on different operating systems, applications, or browsers without losing data. Though it was initially designed to address the issues in publishing voluminous data, XML plays a key role in the exchange and transportation of a huge variety of information and XML documents can be easily accessed by any programming language such as Java etc.

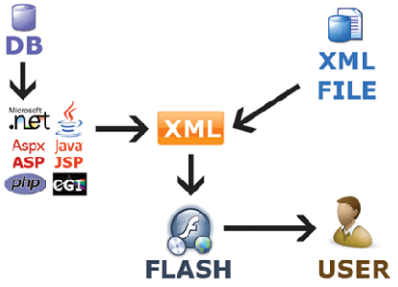


Fig. 1. XML Architecture

APIs (Application Programming Interface) ease the task of accessibility. The potential of XML has greatly been harnessed in the Web with the help of an artifact called web service. In distributed world, the designing of the web services is based on XML technology as the underneath building block. The structure of an XML document is described in XML schema [6] which is used to structure and delimit the allowable values for XML documents. XML has emerged as an efficient medium for information transfer and attempting to offer similar kind of information environment up to an extent. For better understanding to the reader, a sample example of XML is shown in figure 2.

```
<?xml version="1.0"?>
<menu>
   <parent title="kirupaPicks">
    <child>
        <title>kirupa.com</title>
        <link>http://www.kirupa.com</link>
    </child>
    <child>
        <title>kirupaForum</title>
        <link>http://www.kirupa.com/forum/</link>
    </child>
    <child>
        <title>kirupa's Blog</title>
        <link>http://blog.kirupa.com</link>
    </child>
   </parent>
</menu>
```

Fig. 2. Sample XML file

Lately, the scientific community has recognized the potential of XML as middle man to provide interoperability support in heterogeneous distributed environment and works as format changer at the middle tier. The format changing is feasible using various technologies which mainly built on XML as the building block such as XSLT [9], SOAP [2], and BIZTALK [15]. In this paper we have emphasized on XSLT and SOAP and left the discussion of BIZTALK to the reader.

3 What Is Interoperability

Interoperability describes the ability to work together to deliver services in a seamless, uniform and efficient manner across multiple organizations and information technology systems. In order to produce fruitful outcomes in distributed environment on a global scale diversified systems need to work together. This property of systems is known as system interoperability. Interoperability allows the system to work with other systems- present or future - without any intensive implementation or restricted access. The exposed interfaces of the system are completely understood at the receiving destinations. Numerous definitions of system interoperability have been proposed in literature. According to IEEE glossaries [6], it is the ability of systems (more than two) that exchange information and use it. James et al. defines interoperability as "be able to accomplish end user applications using different types of computer systems, operating systems and application software, interconnected by different types of local and wide area networks." Data formats and communication protocols are fundamental artifacts helpful to archive interoperability. We attempt to draw interoperability in figure 3.

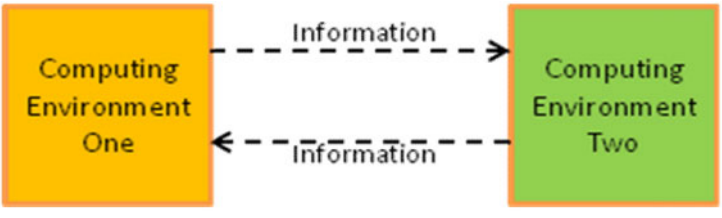


Fig. 3. Graphical representation of interoperability

3.1 Interoperability Analysis

Interoperability is analyzed from following different angles:

- **Organizational interoperability** is associated with the activities of organizations and agreements between them. Organizational interoperability is ensured by legislation and general agreements.
- **Semantic interoperability** refers to the ability of different organizations to understand the exchanged data in a similar way. This presumes the creation of a mechanism allowing the presentation of service data and data definitions.
- **Technical Infrastructure interoperability** is the ability of hardware acquired by different organizations to work in a connected way.
- **Software interoperability** refers to the ability of software used in different organizations to exchange data. Achieving software interoperability requires the establishment of common data exchange protocols, development of software necessary for the management of data connections, and creation of user interfaces in order to enable communication between different organizations.

4 Background and Related Work

W3Consortium is well known for accommodating wide range of technology and providing universal standard for their use [1]. As far as XML technology is concerned, W3C explains about XML Data Binding and how XML is used to adapt to the line end conventions of various modern operating systems. In [2] Lakshaman et al. attempt to explain how XSL Transformations (XSLT) are used to transform XML documents into other XML documents, text documents or HTML documents. Carey et al. [5] explain how XML can be used to support Object Relational Data Mapping. Authors also explain how XML is compatible between different type systems in object-oriented programming languages.

5 Innovative Approaches

This section elaborates the evaluation criteria aimed to be exploited in evaluating XML interoperability. XML's interoperability capabilities will be evaluated using the following criteria:

- **Support for Object Relational Data.** XML intervenes in the mapping of object – on the front end and relational data – on the back end. This is an emerging idea is gaining rapid attention in software application development industry. It is popularly known as ORM [15] and there are various frameworks available today in the commercial, research and literature arenas, which have successfully accommodated ORM. One popular example is Hibernate [15].
- **Intactness of Existing Systems.** Intactness of the system indicates that system environment does not change, post operation on it, if any. In this paper we try to explain, that XML interoperability helps to maintain the systems properties and environments unchanged, if they (systems) exposed to distributed environment - constituted by disparate systems - to coordinate to each other to execute any task.
- **Heterogeneity of Spatial Validity.** By spatial heterogeneity, we mean that, in a distributed computing environment, a single server receives requests from various corners of the globe, which do not share the same language. Good example may be disjoint set of countries like Germany, China, France etc, where a computer user expects the request's output in country national language different from English. In the following section we validate that, XML interoperability helps to realize such dream.
- **Information Exchange via Middle Tier Format Changers.** In distribute environment – containing various disparate computing nodes, all the computing nodes need not to be have similar computing environment. In order to exchange the information in a coordinating way, there need to be a middle tier information format changer, which receives the information from one computing node in one format, converts into another format and propagates to another computing node.

6 Innovation of System Interoperability in XML

This section we broadly discuss whether a particular evaluation criteria satisfies XML interoperability or not and the rationale behind either outcome.

6.1 Supports for Object Relational Data

Object Relational Data Mapping is a programming technique for converting data between incompatible type systems in object-oriented programming languages. XML interoperability plays a key important role in mapping in a mapping file. The class objects at front end (Java object) maps with the respective relational object and back end (table name in database) and helps in data processing between frontend and backend. Java Hibernate works on this model, which relies on mapping file, which is an xml file. The figure 4 depicts the above stated scenario of object relational mapping. This is in accordance to hibernate [15] frame work, widely used at the middleware in software development. The very left side of the figure is a user defined java object called user.java, constituted by username and password. Whereas the extreme right side of the figure is a relational table (regardless of database name) called TB_USER, which has two attributes: USERNAME, PASSWORD. In the middle of these two objects there is a mapping artifact (we call it User.hbm) - hibernate mapping- (XML format) which maps the object property on the client side to table property at the back end and these two objects itself. In other words we can say that XML formatted hbm file at the middleware takes the java request converts into table mapping which is adaptable to the back end and retrieves the result.

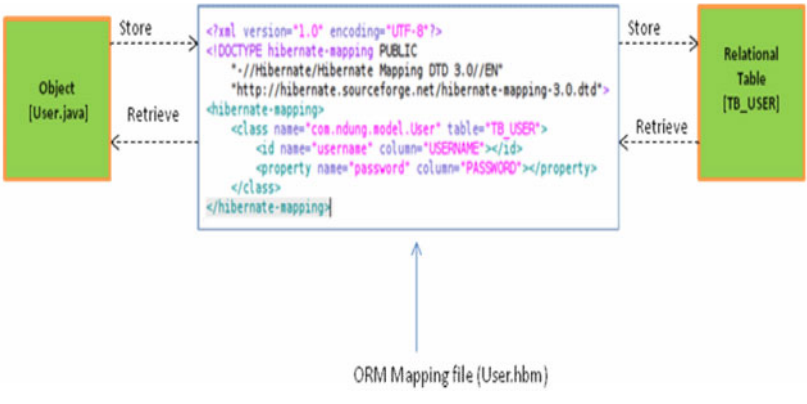


Fig. 4. Object Relational Mapping

6.2 Intactness of Existing Systems

Due scalability of internet access distributed environment is emerging as main stream abode. Web services are playing as technological pillars for the sustainability of distributed environment. Web services are largely dependent on XML in terms of SOAP (Simple Object Access Protocol). SOAP messages which are based on XML

possess the potential for transformation of the way the distributed application written and how the data gets exchanged. SOAP is platform and language neutral and does not depend on any object model. This boasts SOAP to span the SOAP enabled applications into multiple distributed operating systems. So the computer systems environment needs not to be changed at all rather SOAP message takes care in terms of interoperability. A sample SOAP message has been shown in figure 5 to make more understanding to reader.

```
-----SoapRequest at 8/8/2009 10:47:43 PM
<?xml version="1.0" encoding="utf-8"?><soap:Envelope
xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xsd="http://www.w3.org/2001/XMLSchema"><soap:Body><HelloWorld
xmlns="http://tempuri.org/"><myString>This is a test message</myString>
</HelloWorld></soap:Body></soap:Envelope>

-----SoapResponse at 8/8/2009 10:47:45 PM
<?xml version="1.0" encoding="utf-8"?><soap:Envelope
xmlns:soap="http://schemas.xmlsoap.org/soap/envelope/"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:xsd="http://www.w3.org/2001/XMLSchema"><soap:Body>
<HelloWorldResponse xmlns="http://tempuri.org/"><HelloWorldResult>Hello
World Service returns - This is a test message</HelloWorldResult>
</HelloWorldResponse></soap:Body></soap:Envelope>
```

Fig. 5. Sample SOAP Message

XML is a platform independent language in order to operate it we need to change existing environment of the system. Suppose there is a computer system running on windows environment and there is another one running on Linux environment there may be possibility that some additional API's (Application Programming Interface) might need it such as SOAP. Few web services may be might be needed but XML's interoperability makes possible to open this document on both the environments. Hence by changing the SOAP envelope suitable.

6.3 Heterogeneity of Spatial Validity

XML's interoperability is emerging as a divine to cater the need of information exchange among systems located at various corner in the world which rely on their own specific language need not to be similar to other language. The figure 6 explains the scenario. It shows a client server architecture where server is enriched with XSTL (Extensible Stylesheet Language Transformations) [9] capabilities. The environment is distributed and server receives request from various parts of the world, four countries as shown in figure. The request encapsulates country specific language information which server harnesses to transform that XML document - from the database - into appropriate language at the client terminal. Thus XML supports Heterogeneity of Spatial Validity.

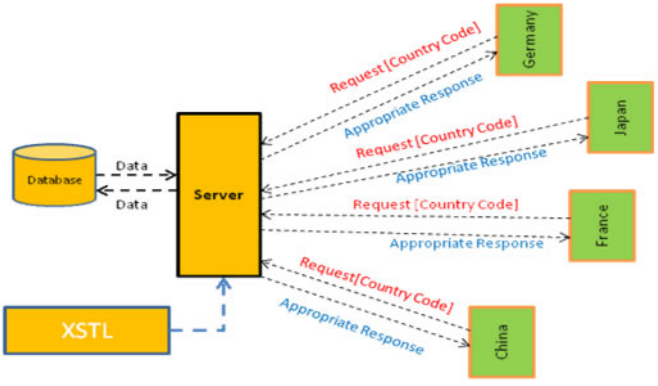


Fig. 6. Spatial Validity Demonstration

6.4 Information Exchange via Middle Tier Format Changers

Middle Tier Exchangers- a third party framework exchanger possess the functionality to accept inputs in any formats converts into equivalent XML format and send output as desired output. Thus underneath the technology the XML interoperability serves as building block. Hence information can be exchanged regardless computing environment. As depicted in the figure 7 we consider computer environment two is database environment (spatio-temporal) where spatial artifact is in GML [10] format which is nothing but XML format. Computing environment one is Geographic Environment System (GIS) [12]. When GIS community (format y) needs to exploit the GML data from computing environment one (format x), the format needs to be changed into the GIS specific format which is Shape file format. There needs to be format exchanger which converts the GML format into shape format. We assume that a customized third party tool [11] is available which relies on XML format for its metadata for data conversion. Thus the third party format changer is placed logically between both the environments to facilitate the format changing process.

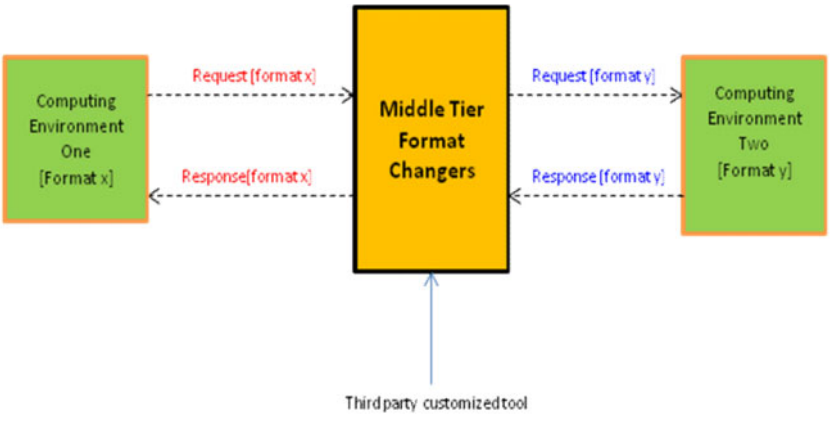


Fig. 7. Middle Tier Format Changer for XML

7 Conclusion

Interoperability tends to be regarded as an issue for experts and its implications for daily living are sometimes underrated. So many technologies and organizations are dedicated to interoperability. Interoperability helps to users get the most out of technology, and it also encourages innovation in the industrial sphere. In this paper we explored XML interoperability based on quite a few hypothetical evaluation criteria. We employed five criteria and evaluated whether XML interoperability supports it or not. In all evaluation approaches we observed that XML worked as a middleware in different flavors such as XSLT, hbm (hibernate mapping file) etc. We expect this research work will be a welcome contribution to the literature and opens avenues for practitioners and researchers in future.

References

[1] W3C. Extensible markup language (xml) 1.0, w3c recommendation. Technical report, W3C
[2] Lakshaman, L.V.S., Sadri, F.: XML interoperability, http://www.uncg.edu/~sadrif/papers/full-xmlinterop.pdf
[3] Bosak, J., Bray, T.: XML and the second-generation web. Scientific American (May 1999)
[4] St. Laurent, S.: Object Developers Group - XML SIG (June 2000)
[5] Carey, M., Florescu, D.: Publishing Object-Relational Data as XML, http://www.cs.cornell.edu/people/jai/papers/XperantoOverview.pdf
[6] http://www.w3schools.com/Schema/schema_intro.asp
[7] Institute of Electrical and Electronics Engineers. IEEE Standard Computer Dictionary: A Compilation of IEEE Standard Computer Glossaries, New York, NY (1990)
[8] O'Brien, J., Marakas, G.: Introduction to Information Systems, 13th edn. McGraw-Hill, ISBN 0073043559
[9] http://en.wikipedia.org/wiki/XSLT
[10] Sharma, S., Gadia, S.: Perl Status Reporter (SRr) on Spatiotemporal Data Mining. International Journal Computer Science and Engineering Survey, AIRCC- IJCSES (August 2010)
[11] http://www.deegree.org
[12] Loesgen, B.: XML Interoperability, Abrufam (February 17, 2003), http://www.vbxml.com/conference/wrox/2000_vegas/Powerpoints/brianl_xml.pdf
[13] Pohl, K., Bockle, G., van der Linden, F.: Software Product Line Engineering: Foundations, Principles and Technique. Springer (July 2005)
[14] Wohlin, C., Ahlgren, M.: Soft factors and their impact on time to market. Software Quality Journal 4, 189–205 (1995)
[15] Sundmaeker, H., Guillemin, P., Friess, P., Woelfflé, S.: Vision and Challenges for Realising the Internet of Things. In: Cluster of European Research Projects on the Internet of Things (CERP-IoT) (2010)
[16] Bray, T., Paoli, J., Sperberg-McQueen, C.M., et al.: Extensible Markup Language (XML), 3rd edn., http://www.w3.org/TR/REC-xml

Author Index