

Yo-Sung Ho (Ed.)

LNCIS 7088

Advances in Image and Video Technology

5th Pacific Rim Symposium, PSIVT 2011
Gwangju, South Korea, November 2011
Proceedings, Part II

2
Part II



Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany

Yo-Sung Ho (Ed.)

Advances in Image and Video Technology

5th Pacific Rim Symposium, PSIVT 2011
Gwangju, South Korea, November 20-23, 2011
Proceedings, Part II

Volume Editor

Yo-Sung Ho

Gwangju Institute of Science and Technology (GIST)
1 Oryong-dong Buk-gu, Gwangju, 500-712, South Korea
E-mail: hoyo@gist.ac.kr

ISSN 0302-9743

e-ISSN 1611-3349

ISBN 978-3-642-25345-4

e-ISBN 978-3-642-25346-1

DOI 10.1007/978-3-642-25346-1

Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011940679

CR Subject Classification (1998): H.5.1, H.5, I.4-5, I.2.10, I.3, H.3-4, E.4

LNCS Sublibrary: SL 6 – Image Processing, Computer Vision, Pattern Recognition,
and Graphics

© Springer-Verlag Berlin Heidelberg 2011

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper

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

Preface

We are delighted to welcome readers to the proceedings of the 5th Pacific-Rim Symposium on Video and Image Technology (PSIVT 2011), held in Gwangju, Korea, during November 20-23, 2011. The first PSIVT was held in Hsinchu, Taiwan, in 2006. Since then, it has been hosted successfully by Santiago, Chile, in 2007, Tokyo, Japan, in 2009, Singapore in 2010, and finally Gwangju, one of the beautiful and democratic cities in Korea. The symposium provides a forum for presenting and discussing the latest research and development in image and video technology and explores possibilities and future directions in the field. PSIVT 2011 continued to attract researchers, artists, developers, educators, performers, and practitioners of image and video technology from the Pacific rim and around the world.

In PSIVT 2011, the Program Committee was made up of Area Chairs and a Technical Program Committee. The technical areas of PSIVT 2011 covered Image/Video Coding and Transmission, Image/Video Processing and Analysis, Imaging and Graphics Hardware and Visualization, Image/Video Retrieval and Scene Understanding, Biomedical Image Processing and Analysis, Biometrics and Image Forensics, and Computer Vision Applications. For each technical area, at least two Area Chairs were assigned to coordinate the paper-review process with their own team of reviewers selected from the Technical Program Committee. The review process was double-blind in which author names and affiliations were not made known to Area Chairs and reviewers. Reviewers also did not know their Area Chairs. Each paper received at least three reviews. The reviewers were asked to submit a detailed review report and the Area Chairs made the final decisions on the acceptance of papers with little moderation from the Program Chairs. In PSIVT 2011, we accepted 71 papers out of 168 submissions including oral and poster session papers. The acceptance rate of 42% indicates our commitment to ensuring a very high-quality symposium.

PSIVT 2011 was organized by the Realistic Broadcasting Research Center (RBRC) at Gwangju Institute of Science and Technology (GIST) in Korea. The symposium was supported by the Center for Information Technology Education (BK21) at GIST, Gwangju Convention and Visitors Bureau, and the MPEG Forum in Korea.

This symposium would not be possible without the efforts of many people. First of all, we are very grateful to all the authors who contributed their high-quality research work and shared their knowledge with our scientific community. We would also like to appreciate the full support of the excellent Program

Committee and all reviewers that provided timely and insightful reviews. Finally, our thanks must go to all members of the Organizing and Steering Committee for their precious time and enthusiasm. They did their best in financing, publicity, publication, registration, Web and local arrangements.

November 2011

Yo-Sung Ho

PSIVT 2011 Organization

Organizing Committee

General Co-chairs

Yo-Sung Ho	Gwangju Institute of Science and Technology, Korea
Wen-Nung Lie	National Chung Cheng University, Taiwan
Domingo Mery	Pontificia Universidad Catolica, Chile

Program Co-chairs

Kap Luk Chan	Nanyang Technological University, Singapore
Qingming Huang	Chinese Academy of Sciences, China
Shin'ichi Satoh	National Institute of Informatics, Japan

Finance Chair

Kuk-Jin Yoon	Gwangju Institute of Science and Technology, Korea
--------------	---

Publicity Co-chairs

Sung-Hee Lee	Gwangju Institute of Science and Technology, Korea
Yousun Kang	Tokyo Polytechnic University, Japan

Publication Chair

Sung Chan Jun	Gwangju Institute of Science and Technology, Korea
---------------	---

Local Arrangements Chair

Hyunju Lee	Gwangju Institute of Science and Technology, Korea
------------	---

Steering Committee

Kap Luk Chan	Nanyang Technological University, Singapore
Yung-Chang Chen	National Tsinghua University, Taiwan
Yo-Sung Ho	Gwangju Institute of Science and Technology, Korea
Reinhard Klette	The University of Auckland, New Zealand
Wen-Nung Lie	National Chung Cheng University, Taiwan
Domingo Mery	Pontificia Universidad Catolica, Chile
Akihiro Sugimoto	National Institute of Informatics, Japan
Mohan M. Trivedi	University of California, San Diego, USA

Area Chairs

Oscar Au	Hong Kong University of Science and Technology, Hong Kong
Miguel Carrasco	Universidad Diego Portales, Chile
Yoong Choon Chang	Multimedia University, Malaysia
Anthony TS Ho	University of Surrey, UK
Fay Huang	National Ilan University, Taiwan
Shuaqiang Jiang	Chinese Academy of Sciences, China
Shang-Hong Lai	National Tsing Hua University, Taiwan
Jaejoon Lee	Samsung Electronics, Korea
Qingshan Liu	Rutgers University, USA
Chia-Wen Lin	National Tsing-Hua University, Taiwan
Huei-Yung Lin	National Chung Cheng University, Taiwan
Yasuhiro Mukaigawa	Osaka University, Japan
Luis Pizarro	Imperial College, UK
Mingli Song	Zhejiang University, China
Yu-Wing Tai	KAIST, Korea
Gang Wang	Nanyang Technological University, Singapore
Lei Wang	University of Wollongong, Australia
Changsheng Xu	Chinese Academy of Sciences, China
Shuicheng Yan	National University of Singapore, Singapore
Junsong Yuan	Nanyang Technological University, Singapore
Jianxin Wu	Nanyang Technological University, Singapore
Vitali Zagorodnov	Nanyang Technological University, Singapore

Technical Program Committee

Hezerul Abdul Karim	Michael Cree
Toshiyuki Amano	Ismael Daribo
Yasuo Ariki	Xiaoyu Deng
Vishnu Monn Baskaran	Lei Ding
Bedrich Benes	Zhao Dong
Xiujuan Chai	Gianfranco Doretto
Yoong Choon Chang	How-Lung Eng
Chin-Chen Chang	Giovani Gomez
Chia-Yen Chen	Gerardo Fernández-Escribano
Yi-Ling Chen	Chiou-Shann Fuh
Chu-Song Chen	Makoto Fujimura
Jia Chen	Hironobu Fujyoshi
Hwann-Tzong Chen	Kazuhiro Fukui
Jian Cheng	Simon Hermann
Gene Cheung	Yo-Sung Ho
Chen-Kuo Chiang	Seiji Hotta
Sunghyun Cho	Jun-Wei Hsieh

Changbo Hu	Takahiro Okabe
Xiaoqin Huang	Ho-Yuen Pang
Rui Huang	Christian Pieringer
Junzhou Huang	Lei Qin
Chun-Rong Huang	Bo Qiu
Naoyuki Ichimura	Mauricio Reyes
Masahiro Iwahashi	Laurent Risser
Daisuke Iwai	Isaac Rudomin
Yoshio Iwai	Clarisa Sanchez
Gangyi Jiang	Tomokazu Sato
Xin Jin	Takeshi Shakunaga
Ramakrishna Kakarala	Shiguang Shan
Masayuki Kanbara	Xiaowei Shao
Li-Wei Kang	Chunhua Shen
Hiroshi Kawasaki	Ikuko Shimizu
Chang-Su Kim	Keita Takahashi
Itaru Kitahara	Toru Tamaki
Mario Koeppen	Ping Tan
Akira Kubota	Masayuki Tanaka
Takio Kurita	Flavio Torres
Shang-Hong Lai	Chien-Cheng Tseng
Tung-Ying Lee	Seiichi Uchida
Wen-Nung Lie	Carlos Vazquez
Chia-Wen Lin	Yu-Chiang Wang
Guo-Shiang Lin	Jingqiao Wang
Xiao Liu	Min-Liang Wang
Damon Shing-Min Liu	Hsien-Huang Wu
Huiying Liu	Ming Yang
Jonathan Loo	Chia-Hung Yeh
Yasushi Makihara	Kaori Yoshida
Takeshi Masuda	Guangtao Zhai
Fabrice Meriadeau	Daoqiang Zhang
Rodrigo Moreno	Qi Zhao
Hajime Nagahara	Yuanjie Zheng
Atsushi Nakazawa	Bo Zheng
Kai Ni	Huiyu Zhou
Shohei Nobuhara	Shaohua Zhou
Takeshi Oishi	

Sponsoring Institutions

The Realistic Broadcasting Research Center (RBRC) at GIST
The Center for Information Technology Education (BK21) at GIST
Gwangju Convention and Visitors Bureau
The MPEG Forum in Korea

Table of Contents – Part II

Lossless Image Coding Based on Inter-color Prediction for Ultra High Definition Image	1
<i>Jiho Park, Je-Woo Kim, Jechang Jeong, and Byeongho Choi</i>	
Multithreading Architecture for Real-Time MPEG-4 AVC/H.264 SVC Decoder	13
<i>Yong-Hwan Kim, Jiho Park, and Je-Woo Kim</i>	
Fast Mode Decision Algorithm for Depth Coding in 3D Video Systems Using H.264/AVC	25
<i>Da-Hyun Yoon and Yo-Sung Ho</i>	
Improved Diffusion Basis Functions Fitting and Metric Distance for Brain Axon Fiber Estimation	36
<i>Ramón Aranda, Mariano Rivera, and Alonso Ramírez-Manzanares</i>	
An Adaptive Motion Data Storage Reduction Method for Temporal Predictor	48
<i>Ruobing Zou, Oscar C. Au, Lin Sun, Sijin Li, and Wei Dai</i>	
A Local Variance-Based Bilateral Filtering for Artifact-Free Detail- and Edge-Preserving Smoothing	60
<i>Cuong Cao Pham, Synh Viet Uyen Ha, and Jae Wook Jeon</i>	
Iterative Gradient-Driven Patch-Based Inpainting	71
<i>Sarawut Tae-o-sot and Akinori Nishihara</i>	
Feature Extraction Based on Co-occurrence of Adjacent Local Binary Patterns	82
<i>Ryusuke Nosaka, Yasuhiro Ohkawa, and Kazuhiro Fukui</i>	
Natural Image Composition with Inhomogeneous Boundaries	92
<i>Dong Wang, Weijia Jia, Guiqing Li, and Yunhui Xiong</i>	
Directional Eigentemplate Learning for Sparse Template Tracker	104
<i>Hiroyuki Seto, Tomoyuki Taguchi, and Takeshi Shakunaga</i>	
Gender Identification Using Feature Patch-Based Bayesian Classifier	116
<i>Shen-Ju Lin, Chung-Lin Huang, and Shih-Chung Hsu</i>	
Multiple Objects Tracking across Multiple Non-overlapped Views	128
<i>Ke-Yin Chen, Chung-Lin Huang, Shih-Chung Hsu, and I-Cheng Chang</i>	

Fast Hypercomplex Polar Fourier Analysis for Image Processing	141
<i>Zhuo Yang and Sei-ichiro Kamata</i>	
Colorization by Landmark Pixels Extraction	149
<i>Weiwei Du, Shiya Mori, and Nobuyuki Nakamori</i>	
Filtering-Based Noise Estimation for Denoising the Image Degraded by Gaussian Noise	157
<i>Tuan-Anh Nguyen and Min-Cheol Hong</i>	
Combining Mendonça-Cipolla Self-calibration and Scene Constraints . . .	168
<i>Adlane Habed, Tarik Elamsy, and Boubakeur Boufama</i>	
A Key Derivation Scheme for Hierarchical Access Control to JPEG 2000 Coded Images	180
<i>Shoko Imaizumi, Masaaki Fujiyoshi, Hitoshi Kiya, Naokazu Aoki, and Hiroyuki Kobayashi</i>	
Bifocal Matching Using Multiple Geometrical Solutions	192
<i>Miguel Carrasco and Domingo Mery</i>	
Digital Hologram Compression Using Correlation of Reconstructed Object Images	204
<i>Jae-Young Sim</i>	
Pedestrian Image Segmentation via Shape-Prior Constrained Random Walks	215
<i>Ke-Chun Li, Hong-Ren Su, and Shang-Hong Lai</i>	
A Novel Rate Control Algorithm for H.264/AVC Based on Human Visual System	227
<i>Jiangying Zhu, Mei Yu, Qiaoyan Zheng, Zongju Peng, Feng Shao, Fucui Li, and Gangyi Jiang</i>	
Blind Image Deblurring with Modified Richardson-Lucy Deconvolution for Ringing Artifact Suppression	240
<i>Hao-Liang Yang, Yen-Hao Chiao, Po-Hao Huang, and Shang-Hong Lai</i>	
Quality Estimation for H.264/SVC Inter-layer Residual Prediction in Spatial Scalability	252
<i>Ren-Jie Wang, Yan-Ting Jiang, Jiunn-Tsair Fang, and Pao-Chi Chang</i>	
Extracting Interval Distribution of Human Interactions	262
<i>Ryohei Kimura, Noriko Takemura, Yoshio Iwai, and Kosuke Sato</i>	

A Flexible Method for Localisation and Classification of Footprints of Small Species	274
<i>Haokun Geng, James Russell, Bok-Suk Shin, Radu Nicolescu, and Reinhard Klette</i>	
Learning and Regularizing Motion Models for Enhancing Particle Filter-Based Target Tracking	287
<i>Francisco Madrigal, Mariano Rivera, and Jean-Bernard Hayet</i>	
CT-MR Image Registration in 3D K-Space Based on Fourier Moment Matching	299
<i>Hong-Ren Su and Shang-Hong Lai</i>	
Sparse Temporal Representations for Facial Expression Recognition	311
<i>S. W. Chew, R. Rana, P. Lucey, S. Lucey, and S. Sridharan</i>	
Dynamic Compression of Curve-Based Point Cloud	323
<i>Ismael Daribo, Ryo Furukawa, Ryusuke Sagawa, Hiroshi Kawasaki, Shinsaku Hiura, and Naoki Asada</i>	
Recovering Depth Map from Video with Moving Objects	335
<i>Hsiao-Wei Chen and Shang-Hong Lai</i>	
An Iterative Algorithm for Efficient Adaptive GOP Size in Transform Domain Wyner-Ziv Video Coding	347
<i>Khanh Dinh Quoc, Xiem Hoang Van, and Byeungwoo Jeon</i>	
A Robust Zero-Watermark Copyright Protection Scheme Based on DWT and Image Normalization	359
<i>Mahsa Shakeri and Mansour Jamzad</i>	
Multi-view Video Coding Based on High Efficiency Video Coding	371
<i>Kwan-Jung Oh, Jaejoon Lee, and Du-Sik Park</i>	
2D to 3D Image Conversion Based on Classification of Background Depth Profiles	381
<i>Guo-Shiang Lin, Han-Wen Liu, Wei-Chih Chen, Wen-Nung Lie, and Sheng-Yen Huang</i>	
Shape Matching and Recognition Using Group-Wised Points	393
<i>Junwei Wang, Yu Zhou, Xiang Bai, and Wenyu Liu</i>	
Author Index	405

Table of Contents – Part I

Nonlinear Transfer Function-Based Image Detail Preserving Dynamic Range Compression for Color Image Enhancement	1
<i>Deepak Ghimire and Joonwhoan Lee</i>	
3D Perception Adjustment of Stereoscopic Images Based upon Depth Map	13
<i>Jong In Gil, Seung Eun Jang, and Manbae Kim</i>	
Super-Resolved Free-Viewpoint Image Synthesis Using Semi-global Depth Estimation and Depth-Reliability-Based Regularization	22
<i>Keita Takahashi and Takeshi Naemura</i>	
Heat Kernel Smoothing via Laplace-Beltrami Eigenfunctions and Its Application to Subcortical Structure Modeling	36
<i>Seung-Goo Kim, Moo K. Chung, Seongho Seo, Stacey M. Schaefer, Carien M. van Reekum, and Richard J. Davidson</i>	
SLAM and Navigation in Indoor Environments	48
<i>Shang-Yen Lin and Yung-Chang Chen</i>	
Color Based Stool Region Detection in Colonoscopy Videos for Quality Measurements	61
<i>Jayantha Muthukudage, JungHwan Oh, Wallapak Tavanapong, Johnny Wong, and Piet C. de Groen</i>	
Improving Motion Estimation Using Image-Driven Functions and Hybrid Scheme	73
<i>Duc Dung Nguyen and Jae Wook Jeon</i>	
Real-Time Background Compensation for PTZ Cameras Using GPU Accelerated and Range-Limited Genetic Algorithm Search	85
<i>Thuy Tuong Nguyen and Jae Wook Jeon</i>	
Audio-Visual Speech Recognition Based on AAM Parameter and Phoneme Analysis of Visual Feature	97
<i>Yuto Komai, Yasuo Arika, and Tetsuya Takiguchi</i>	
Multi-scale Integration of Slope Data on an Irregular Mesh	109
<i>Rafael F.V. Saracchini, Jorge Stolfi, Helena C.G. Leitão, Gary Atkinson, and Melvyn L. Smith</i>	
Virtual Viewpoint Disparity Estimation and Convergence Check for Real-Time View Synthesis	121
<i>In-Yong Shin and Yo-Sung Ho</i>	

Spatial Feature Interdependence Matrix (SFIM): A Robust Descriptor for Face Recognition	132
<i>Anbang Yao and Shan Yu</i>	
Coding of Dynamic 3D Mesh Model for 3D Video Transmission	144
<i>Jui-Chiu Chiang, Chun-Hung Chen, and Wen-Nung Lie</i>	
Ray Divergence-Based Bundle Adjustment Conditioning for Multi-view Stereo	153
<i>Mauricio Hess-Flores, Daniel Knoblauch, Mark A. Duchaineau, Kenneth I. Joy, and Falko Kuester</i>	
Temporally Consistent Disparity and Optical Flow via Efficient Spatio-temporal Filtering	165
<i>Asmaa Hosni, Christoph Rhemann, Michael Bleyer, and Margrit Gelautz</i>	
Specular-Free Residual Minimization for Photometric Stereo with Unknown Light Sources	178
<i>Tsuyoshi Migita, Kazuhiro Sogawa, and Takeshi Shakunaga</i>	
Analysing False Positives and 3D Structure to Create Intelligent Thresholding and Weighting Functions for SIFT Features	190
<i>Michael May, Martin Turner, and Tim Morris</i>	
Verging Axis Stereophotogrammetry	202
<i>Khurram Jawed and John Morris</i>	
More on Weak Feature: Self-correlate Histogram Distances	214
<i>Sheng Wang, Qiang Wu, Xiangjian He, and Wenjing Jia</i>	
Mid-level Segmentation and Segment Tracking for Long-Range Stereo Analysis	224
<i>Simon Hermann, Anko Börner, and Reinhard Klette</i>	
Applications of Epsilon Radial Networks in Neuroimage Analyses	236
<i>Nagesh Adluru, Moo K. Chung, Nicholas T. Lange, Janet E. Lainhart, and Andrew L. Alexander</i>	
Road Image Segmentation and Recognition Using Hierarchical Bag-of-Textons Method	248
<i>Yusun Kang, Koichiro Yamaguchi, Takashi Naito, and Yoshiki Ninomiya</i>	
On the Security of a Hybrid SVD-DCT Watermarking Method Based on LPSNR	257
<i>Huo-Chong Ling, Raphael C.-W. Phan, and Swee-Huay Heng</i>	

Improved Entropy Coder in H.264/AVC for Lossless Residual Coding in the Spatial Domain	267
<i>Jin Heo and Yo-Sung Ho</i>	
Attention Prediction in Egocentric Video Using Motion and Visual Saliency	277
<i>Kentaro Yamada, Yusuke Sugano, Takahiro Okabe, Yoichi Sato, Akihiro Sugimoto, and Kazuo Hiraki</i>	
FAW for Multi-exposure Fusion Features	289
<i>Michael May, Martin Turner, and Tim Morris</i>	
Efficient Stereo Image Rectification Method Using Horizontal Baseline	301
<i>Yun-Suk Kang and Yo-Sung Ho</i>	
Real-Time Image Mosaicing Using Non-rigid Registration	311
<i>Rafael Henrique Castanheira de Souza, Masatoshi Okutomi, and Akihiko Torii</i>	
Adaptive Guided Image Filtering for Sharpness Enhancement and Noise Reduction	323
<i>Cuong Cao Pham, Synh Viet Uyen Ha, and Jae Wook Jeon</i>	
Half-Sweep Imaging for Depth from Defocus	335
<i>Shuhei Matsui, Hajime Nagahara, and Rin-ichiro Taniguchi</i>	
A Hierarchical Approach to Practical Beverage Package Recognition	348
<i>Mei-Chen Yeh and Jason Tai</i>	
An Equivalent 3D Otsu’s Thresholding Method	358
<i>Puthipong Sthitpattanapongsa and Thitiwan Srinark</i>	
Human Motion Tracking with Monocular Video by Introducing a Graph Structure into Gaussian Process Dynamical Models	370
<i>Jianfeng Xu, Koichi Takagi, and Shigeyuki Sakazawa</i>	
Depth Map Up-Sampling Using Random Walk	384
<i>Gyo-Yoon Lee and Yo-Sung Ho</i>	
Evaluation of a New Coarse-to-Fine Strategy for Fast Semi-Global Stereo Matching	395
<i>Simon Hermann and Reinhard Klette</i>	
Theoretical Analysis of Multi-view Camera Arrangement and Light-Field Super-Resolution	407
<i>Ryo Nakashima, Keita Takahashi, and Takeshi Naemura</i>	
Author Index	421

Lossless Image Coding Based on Inter-color Prediction for Ultra High Definition Image

Jiho Park¹, Je Woo Kim¹, Jechang Jeong², and Byeongho Choi¹

¹ Multimedia IP Research Center, KETI,

25 Saenari-ro Yatap-dong, Bundang-gu, Seongnam-si, Gyeonggi-do 463816, Korea

² Department of Electronic and Computer Engineering, Hanyang University,
222, Wangsimni-ro, Seongdong-gu, Seoul 133791, Korea

{scottie, jwkim, bhchoi}@keti.re.kr

jjeong@ece.hanyang.ac.kr

Abstract. This paper addresses the lossless image coding for ultra-high definition television system which supports 4K (4096×2160) resolution image with 22.2ch audio. Ultra High Definition Tele-vision system is being developed to satisfy end-user who has a longing for higher resolution, higher quality, and higher fidelity picture and sound. However, the major characteristic of Ultra High Definition Tele-vision system is considerably huge input information must be processed in real-time compare to conventional systems. Therefore high speed data handling for editing and playing, high speed signal interface between devices and real-time source codecs without delay for saving memory space are unavoidable requirements of ultra-high system. This paper focused on lossless image codec for the reason of real-time processing with reasonable coding gain and the proposed algorithms is pixel based algorithms called spatio-color prediction. The proposed algorithm uses inter-color correlation with spatial correlation appropriately and it shows 5.5% coding efficiency improvement compare to JPEG-LS. The simulation was performed using 4K resolution RGB 10bit images and it is claimed that the proposed lossless coding was verified under the developed Ultra High Definition Tele-vision system.

Keywords: lossless coding, ultra high definition, high-fidelity, inter-color.

1 Introduction

Nowadays, many multimedia products are introduced on consumer electronics market and consumers are looking for high speed electronic devices with both high resolution and better quality more and more. To meet the high demand at the market, a lot of products supporting high quality, high speed, high resolution and high fidelity are being developed by broadcasting equipment industry and professional video equipment industry. Analog TV is very rapidly replaced by HDTV which is one of the most familiar products and the end-user expects better performance products. Therefore, the industry needs to develop the next generation product to keep pace with the evolution of environment.

The most predictable and feasible next generation media service after prevailing HDTV devices at home consumer market would be focused on high resolution and better quality image service to be provided. World cinema industries have already introduced the production system using digital cameras, which can provide digital contents with 4K (4096×2160) resolution exceeding 2K (2048×1080). Also on broadcasting field, many researches regarding commercialization of Ultra HDTV have already been materialized evacuating from the current HDTV (1920×1080). Japan has announced in 2008 its implementation plan of next generation broadcasting system to be exploiting from 2015 and UK has a plan to broadcast 2012 London Olympic Games by Ultra High Definition Television (UHDTV).

Meanwhile, the international standardization organizations such as ISO/IEC JTC1 SC29WG11 MPEG and ITU-T SG16 Q.6 VCEG have started international standardization procedures on over 2K resolution videos under the name of HEVC (High Efficiency Video Coding).

Under such a circumstance, UHDTV system is needed badly on every media environmental fields including the technology development of Ultra High Definition (UHD) contents to edit, store, and play freely for fulfillment of the UHDTV system. But huge input information can hardly be edited, stored, and played freely at real time under the current media technology, though the real-time process of input information is imperative on UHDTV system.

Also lossless coding would be essential to satisfy users who want to have high quality resolution without losses. The most suitable compression technology at present might be high speed JPEG-LS based-technology which has efficient compression effects on information processing. But current prevailing compression technology can hardly store the high capacity input information and it is needed to develop high efficiency compression algorithms in order to have an appropriate and efficient UHDTV system.

The developed system is a comprehensive total system supporting UHD image which is delivered from 4K digital camera and displayed by four 2K 10bit LCD monitors. Furthermore the system architecture for high speed data storage and optical transmission for high speed interface are under development, too.

The specification of input images used on UHDTV system is RGB 4:4:4 10bit depth and it has totally different characteristics of images from those used on conventional applications such as YUV 4:2:0 8bit depth. Thus, verified algorithms for 10bit depth image with reasonable coding efficiency are required to save storage space and also no delay and minimum complexity is required to process huge input information at the same time. Accordingly this paper has been prepared to improve the compression algorithm which has RGB 4:4:4 10bit depth 4K image and to provide modified JPEG-LS [1]-[3] based algorithms.

Following sections would be consisted of the explanation of the proposed inter-color prediction on section 2, the new context modeling for 4K image on section 3, the description of used test images on section 4, the simulation results on section 5 and then the conclusion on the section 6.

2 The Proposed Lossless Coding

In this section, the description of newly developed lossless coding is provided. The main concept of the proposed algorithms is exploiting spatial correlation of UHD image because 4K image has stronger spatial correlation compared to less 2K images. To improve coding efficiency, spatio-color prediction scheme is introduced which is pixel based algorithms such as [3]-[6] and it brought the concept of SICLIC [7] for predictions.

2.1 Inter-color Prediction

In this paper, all images introduced are using RGB color space and each pixel has 10bit depth, and they are coded in order of $G(reen) \rightarrow B(lue) \rightarrow R(ed)$ line by line. As G color plane has the strongest correlation among the other color planes, the G color plane has been coded at the first place, and through this prior coding of G color plane, the reconstructed G color plane can be used as a reference by B and R color planes. In this proposed inter-color prediction, all the pixels are coded pixel by pixel and, in addition to that, the same prediction scheme as JPEG-LS is applied to the G color plane.

If inter-color correlation between current pixels and reference pixels is strong enough, a reference pixel can be used as a predictor of the current pixel and inter-color correlation can be measured by correlation coefficient as shown in (1).

$$S(ICR,SR) = \frac{M \sum_{i=a}^d (ICR_i \times SR_i) - \sum_{i=a}^d ICR_i \times \sum_{i=a}^d SR_i}{\sqrt{(M \sum_{i=a}^d ICR_i^2 - (\sum_{i=a}^d ICR_i)^2) \times (M \sum_{i=a}^d SR_i^2 - (\sum_{i=a}^d SR_i)^2)}} \quad (1)$$

In equation (1), M indicates the number of pixels used ($=4$), SR_i indicates the value of neighboring pixels, ICR_i means the value of corresponding reference pixels, and S represents the strength of inter-color correlation. The corresponding pixels and neighboring pixels are located as shown in Fig. 1.

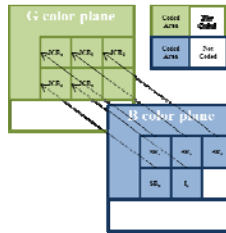


Fig. 1. Location of corresponding reference pixels and neighboring pixels

I_x on the Fig.1 indicates the current encoding/decoding pixel and the correlation coefficient can be obtained through utilizing corresponding 4 pixels from the reference color plane. Needless to say, the value of obtained correlation coefficient is (0~1) and the value indicates the guideline of correlation ratio among corresponding color planes.

Inter-color prediction is implemented when the value of correlation coefficient exceeds 0.5 as strong correlation exists between them and the reference pixels used in encoder side are reconstructed and used identically in decoder side as well.

All the pixels are applied for the inter-color prediction when they have strong correlation exceeding 0.5 and the compensation of color level displacement shall be followed to adjust color intensity in each color plane. If it is assumed that the relationship between ICR_x and I_x is linear, ICR_x can be approximated to I_x ($=I'_x$) and it can be treated as the simple linear equation like $P_{ix}=I'_x=W \times ICR_x+O$ where W is weight factor and O is offset, and P_{ix} is inter-color predictor of I_x . In case of lossless coding, both encoder and decoder can have the value of ICR_x identically, and the value of O is always zero when the floating point operation is possible. Therefore, if the value of W is obtainable, the color level displacement can be compensated. Meanwhile, the value of W can be easily acquired through the minimization of MES as (2).

$$W(ICR,SR)=\frac{M \times \sum_{i=a}^d (ICR_i \times SR_i) - \sum_{i=a}^d ICR_i \times \sum_{i=a}^d SR_i}{\sqrt{\{M \times \sum_{i=a}^d ICR_i^2 - (\sum_{i=a}^d ICR_i)^2\}}} \quad (2)$$

2.2 Spatio-color Prediction

The spatial correlation is very high on 4K image. Generally speaking, 4K image has more correlation for both horizontally and vertically compared to less 2K image when the same object is taken by both 2K camera and 4K camera respectively. It cannot be guaranteed double correlation increasing when the image resolution is doubled by horizontal and vertical direction, but it is certain that 4K image has stronger spatial correlation than less 2K image. Therefore, only when the inter-color correlation is high enough, better coding efficiency improvement can be expected from inter-color prediction than spatial prediction. Furthermore, there is possibility that spatial correlation may exist, though prediction is performed among inter-color. Consequently, a new prediction scheme is needed to be adopted.

According to the spatio-color prediction written in this paper, the improvement of coding efficiency could be achieved from each following difference scheme, after classifying all pixels in B and R color planes into following three categories.

1. If $CC \times PF_1$ exceeds 0.5
2. Else if $CC \times PF_2$ exceeds 0.5
3. Otherwise

The value of PF applied here is a precision factor and it means the accuracy of the correlation to be applied. The value of PF_1 is 0.501 whereas the value of PF_2 is 0.51, which shows the above case 1 is chosen only when the correlation coefficient exceeds 0.998 and above case 2 is chosen, only when the correlation coefficient exceeds 0.98 whereas above case 3 for the balance. As we can realize from the value of the used

threshold, inter-color prediction is applied only when relatively high correlation is kept and coding gain cannot be expected by inter-color prediction from the above case 3. So the same prediction scheme as JPEG-LS is highly recommended in order to prevent coding performance degradation.

In case the current pixel is chosen as the above case 1, unlikely to the conventional method using one predictor either inter-color predictor or spatial predictor, the both are needed for application. For derivation of the new predictor P_{new} , the inter-color predictor can be secured from the corresponding reference pixel ICR_x by multiplying weight factor W with the spatial prediction predictor which can be secured from neighboring pixels avoided an edge. A new spatio-color predictor can be obtained through weighted sum as per following (3) according to characteristics of image.

$$P_{new} = (\alpha)P_{ix} + (1-\alpha)P_x \quad (3)$$

The author has defined the value α , which indicates the characteristics of image, from 0.65 to 0.5 for the purpose of this study. An image like zoom-in has fairly high correlation in spatial, whereas less α value brings better performance result. And high α value brings better performance result, in case of spatially complicated images. In fact, better compression ratio can be expected if we manipulate two values; α and PF , appropriately because the value of α affects reciprocally much to the value of PF , but there remains the possibility of precision problem of floating point operation and that of serious complexity increasing by continuous threshold. During the simulation, general performance results can be achieved when the value of α keeps 0.5.

<p>If spatial prediction direction of ICR_x is horizontal, $P_{ix} = SR_a + W_1 \times (ICR_x - ICR_a)$</p> <p>Else if spatial prediction direction of ICR_x is vertical, $P_{ix} = SR_b + W_1 \times (ICR_x - ICR_b)$</p> <p>Else if spaitl prediction direction of ICR_x is diagonal $P_{ix} = (SR_a + SR_b - SR_c) + W_1 \times (ICR_x - (ICR_a - ICR_b + ICR_c))$</p> <p>Else $P_{ix} = W_2 \times ICR_x$</p>
--

Fig. 2. Derivation process of inter-color prediction for case 2

In case the current pixel is chosen as the above case 2, inter-color correlation is not strong enough like case 1 though weak inter-color prediction scheme is proposed using inter-color correlation as much as possible. Presumption can be made in this proposed method that there is high possibility to have the same directional prediction between the reference color plane and the current color plane. And also residual of neighboring pixels have similar level even though there exists weak inter-color correlation. Therefore P_{ix} can derive from the newly developed algorithm using the prediction direction of reference color plane as per in fig. 2 and final predictor P_{new} can be acquired through (3). For case 2, the recommended value of α is 0.5.

As like case 1, to compensate color level displacement in case 2, the weight factor could be introduced but the coding gain compared to complexity turned out to be negligible because residual of neighboring pixels by spatial prediction are normally small. Accordingly $W_1=1$ is suggested in this paper and W_2 can derive from (2).

In order to implement the proposed method, prediction direction and prediction error value of reference color plane are needed. It has been confirmed through many experiments that the value of prediction error should be remained in a certain area otherwise it would cause performance degradation. Test images used for simulation have 10bit depth therefore 11bit per pixel is needed for residue. However, it is also confirmed in this paper that prediction residue exceeding -15 or 15 does not bring any coding gain, thus the proposed algorithm clipped the absolute value of prediction residual within 0 ~ 15. In case of prediction direction, five directions - skip, horizontal, vertical, diagonal and inter-color prediction are possible to use. Therefore, one byte per pixel was assigned in memory then the prediction residual of reference color plane is stored in LSB 4bits and prediction direction of reference color plane is stored in MSB 4bits.

Moreover, we could find that the reference color plane of B matched with G color plane, whereas both G color plane and B color plane can be reference color plane of R color plane when GBR coded in turn. But indication is needed to decoder on which reference plane has been used for the reference color plane of R color. This indication will bring overhead problem which causes performance degradation in term of the output bitstream size. Both the appointment of reference color plane per line by line and per color plane may be considered to minimize overhead problem. The selective reference color plane method can be employed as an alternative, but serious complexity increasing occurs because the output bitstream would be generated after several iterations by the encoder in order to find the best reference color plane, though decoder would know easily from the parsing of the coded bitstream. For the reason, every reference color plane is restricted within G color plane only in order to minimize encoder complexity increasing and memory usage in this paper.

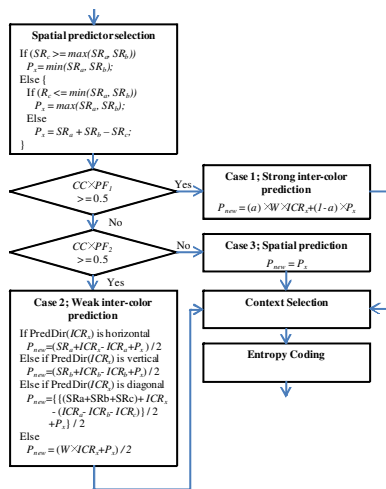


Fig. 3. Flowchart of spatio-color prediction

Finally, in this paper, the same prediction scheme like JPEG-LS is exploited on case 3 which is considered no correlation among the other color planes.

The proposed spatio-color algorithm can be summarized as the flowchart shown in Fig.3

3 Context Modeling

The context model exploited on JPEG-LS can be described by two following steps, calculating the difference of neighboring pixels at first and then quantizing into 9 regions through the thresholds of $[-T_3, -T_2, -T_1, 0, T_1, T_2, T_3]$.

The obtainable differences from 4 neighboring pixels are three kinds as $D_1=SR_d-SR_b$, $D_2=SR_b-SR_c$, $D_3=SR_c-SR_a$ and total 729 contexts ($9 \times 9 \times 9 = 729$) are exploited. However, practically, 364 contexts are employed in regular mode because the same context is used for $[D_1, D_2, D_3]$ and $[-D_1, -D_2, -D_3]$ simultaneously and 2 more contexts are used for run mode additionally.

A new context model is needed for 4K image because the conventional context model does not fully reflect the considerably high spatial correlation. But the prediction scheme of G color plane is the same as conventional method of JPEG-LS so that the context model exploited in G color plane would face a corruption when the conventional model is injured. A new method for maximization of spatial correlation keeping conventional method free from corruption should be invented. So the introduced method in this paper is to classify the average residual of neighboring pixels in to 3 categories at first and then to decide an actual context number of the current pixel after calculating the difference of neighboring pixels as in JPEG-LS as per pseudo code shown in fig. 4.

The proposed method in this paper has been employed under the assumption that spatial correlation has not been perfectly removed, which means the current pixel would have similarity though the direction of spatial prediction is different from that of neighboring pixels. Accordingly, the remained residual might have similar level. For that reason, classifying a similar level residual into the same context would bring improvement of coding efficiency in entropy coding. And the actual context number can be selected within the category as the same method in JPEG-LS because each category has 364 contexts respectively.

- | |
|---|
| <ol style="list-style-type: none"> 1. Calculate the average of $(SR_a \sim SR_d)$ residual
($AVG(SR_a \sim SR_d)$) 2. If $AVG(SR_a \sim SR_d) > 0x08$
Context Number = High context ($732 \leq \text{ContextNum}$) 3. Else if $AVG(SR_a \sim SR_d) > 0x03$
Context Number = Middle context ($367 \leq \text{ContextNum} < 732$) 4. Else
Context Number = Low context ($\text{ContextNum} < 365$) |
|---|

Fig. 4. Flowchart of spatio-color prediction

The context model using the level of residual can be employed in inter-color prediction. As residual is classified by the level of residual, the improved coding performance is expected if some of similar residual can be grouped together even though difference prediction is applied.

4 Test Sequences

The test images used in this paper is from the courtesy of Korea Film Council (KEFIC) [8] taken by RED ONE™ camera of RED Digital Cinema Camera Company [9].

Each image consists of 4096×1716 resolution with RGB 4:4:4 12bit depth and various scenes are taken. The stored file format is DPX and images have been manipulated linear tone mapping to change 12bit depth into 10bit depth for this paper.

The test images which were used in simulation were made to compose similar continuous scenes up to 60 frames and total output bitstream size of whole frames was compared. The comparison of total bitstream size was applied mainly because of the expectation that the developed codec works like high speed real-time video codec in UHDTV system. Thus this research is purposed to measure the coding gain of whole sequence not to measure that of one frame, however during the simulation it is observed that coding gain of similar scene images is not so much fluctuated.

We would like to emphasize that no temporal correlation algorithm has been exploited and more emphasize that each image can be divided separately and can be edited freely as per frame by frame even though encoded at once.

To give better understanding for test images, short description is given as following. Each image is consisted as a set of scenes and each set shows both the first image and the last image respectively.



Fig. 5. The first image of test image sequence 1 (upper) and the last image of test image sequence 1 (lower)

The first sequence as shown in fig.5 shows not much difference between the first image and the last image, but some motions can be found at the movement of big wheel on left side and the movement of windmill on right side.



Fig. 6. The first image of test image sequence 2 (upper) and the last image of test image sequence 2 (lower)

In the second sequence as shown in fig.6, the first image starts with flying helicopter on the left top and ends with scene changes as shown in fig.6 upper image. The fast motion of rollercoaster needs to draw attention on the second sequence. The complexity of image is quite high and motion is also fast.



Fig. 7. The first image of test image sequence 3 (upper) and the last image of test image sequence 3 (lower)

The third sequence as shown in fig.7 shows a parade at an amusement complex and an actress in the center moves very rapidly her arms and the mask on top of stick moves top to down continuously and scene change with zoom-in is occurred as the last image as shown fig.7 lower image. This sequence contains various colors and fast motion of foreground objects.



Fig. 8. The first image of test image sequence 4 (upper) and the last image of test image sequence 4 (lower)

The fourth sequence as shown Fig.8 shows a pair of lovers who sits on the back of a camel and moves forward slowly. The large object moves continuously and slowly and the object moves gradually from bright region to dark region.



Fig. 9. The first image of test image sequence 5 (upper) and the last image of test image sequence 5 (lower)

In the last sequence, the fifth one, there is continuous camera panning on left top direction with showing continuous water streams and water fall.


The tested images used in this paper cover various kinds of scene including almost no movement, scene change, camera panning, small fast object, and large slow object, in which they contains all possible scenes that UHDTV system may encounter.

All images employed are tested under developed UHDTV system, which requires 60 frames per second but only 55 frames were used on the fourth set in order to eliminate scene change.

5 Simulation Results

All test images used for the simulation have 4096×1716 resolution and RGB 4:4:4 10bit depth as described section 4 and consecutive 60 frames are coded excluding fourth test images. In the developed UHDTV system, the maximum supported frame rate will be 60 frames per second and the proposed algorithms shall be verified during 1 second at least. The total size of output bitstream including header syntax was compared to that of JPEG-LS as shown Table. 1. The proposed algorithms are based on lossless image coding, and the reconstruction image is perfectly identical to input original image so that no PSNR comparison is prepared.

Table 1. Font sizes of headings. Table captions should always be positioned *above* the tables

Image	Sequence Number	Total Frame Number	JPEG-LS (Byte)	The Proposed (Byte)	Gain (%)
	Test 1	60 Frames	621,097,483	587,798,614	5.36
	Test 2	60 Frames	579,323,321	546,083,094	5.74
	Test 3	60 Frames	581,731,982	553,109,207	4.92
	Test 4	55 Frames	662,060,078	629,683,219	4.89
	Test 5	60 Frames	784,091,431	732,239,196	6.61
Average					5.50

As shown in Table.1, the proposed algorithms shows up to 6.6%~4.9% coding gain in terms of output bitstream compared to JPEG-LS and 5.5% coding gain can be achieved in average.

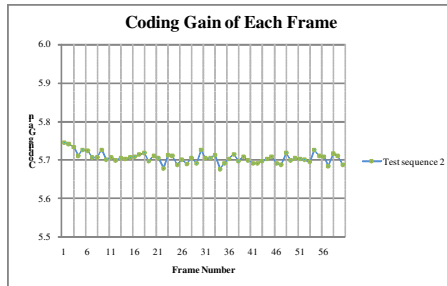


Fig. 10. Coding gain of each frame is compared to JPEG-LS on test sequence 2

In case of the second sequence and the third sequence which contain scene change in the middle of them however, during the simulation, the scene change shows no performance degradation because each frame is coded independently. Fig.10 shows coding gain of each frame compared to JPEG-LS.

6 Conclusion

In this paper, the new lossless algorithm called spatio-color prediction which is very suitable for large image exceeding 2K is proposed and the new context modeling based on residual level is also proposed. The proposed algorithms show improved coding gain up to 6.6% ~ 4.9% in terms of output bitstream size and 5.5% coding gain in average in comparison to JPEG-LS.

The proposed algorithm is verified by the simulation according to various images exploited, which contains various camera actions, foreground movement, and color. Also, the proposed algorithms are adapted in the developed UHDTV system and utilized real-time codec practically.

References

1. ISO/IEC JTC1/SCPS WG1 (JPEG/JBIG): Information Technology - Lossless and Near-Lossless Compression of Continuous-Tone Still Images, Draft International Standard DIS14495-1 (JPEG-LS) (1998)
2. Weinberger, M.J., Seroussi, G., Sapiro, G.: LOCO-I: A Low Complexity, Context-Based, Lossless Image Compression Algorithm. In: Proc. of the 1996 Data Compression Conference, pp. 140–149 (1996)
3. Wu, X.: An Algorithmic Study on Lossless Image Compression. In: Proc. of the 1996 Data Compression Conference, pp. 150–159 (1996)
4. Wu, X.: Lossless Compression of Continuous-Tone Images via Context Selection, Quantization and Modeling. *IEEE Trans. on Image Processing* 6(5), 656–664 (1997)
5. Wu, X., Memon, N., Sayood, K.: A Context-Based, Adaptive, Lossless/Nearly-Lossless Coding Scheme for continuous-Tone Images. ISO/IEC JTC 1/SC 29/WC 1 document No. 202 (1995)
6. Wu, X., Choi, W.K., Memon, N.: Lossless Interframe Image Compression via Context Modeling. In: Proc. of the 1998 Data Compression Conference, pp. 378–387 (1998)
7. Barequet, R., Feder, M.: SICLIC: a simple inter-color lossless image coder. In: Proc. of the 1999 Data Compression Conference, pp. 501–510 (1999)
8. Korea Film Council, <http://koreanfilm.or.kr/>
9. RED Digital Cinema Camera Company, <http://www.red.com/cameras>
10. Matsuda, I., Kaneko, T., Minezawa, A., Itoh, S.: Lossless Coding of Color Images using Block-Adaptive Inter-Color Prediction. In: *IEEE International Conference on Image Processing 2007*, vol. 2, pp. II-329–II-332 (2007)
11. Fukuma, S., Iwahashi, M., Kambayashi, N.: Lossless Color Coordinate Transformer for Lossless Color Image Coding. In: Proc. of IEEE Asia-Pacific Conf. on Circuits and Systems, pp. 595–598 (1998)
12. Golchin, F., Paliwal, K.K.: Minimum-entropy clustering and its application to lossless image coding. In: Proc. of the IEEE International Conference on Image Processing (1997)
13. Memon, N.D., Sayood, K.: Lossless compression of video sequences. *IEEE Trans. Commun.* 44(10), 1340–1345 (1996)

Multithreading Architecture for Real-Time MPEG-4 AVC/H.264 SVC Decoder

Yong-Hwan Kim, Jiho Park, and Je-Woo Kim

Korea Electronics Technology Institute,
Seongnam-si, Gyeonggi-do, Republic of Korea
{yonghwan, scottie, jwkim}@keti.re.kr

Abstract. The inter-layer prediction (ILP) including intra, residual, and motion up-sampling operation in Scalable Video Coding (SVC) significantly increases the compression ratio compared to simulcast. The SVC Codec capable of processing the inter-layer prediction among multiple layers, however, requires much more memory and computational power than single-layer MPEG-4 AVC/H.264 Codec. This paper presents a fast and memory-efficient multithreading architecture for real-time MPEG-4 AVC/H.264 Scalable High profile decoder. Unlike existing approaches where multi-threaded video encoding and decoding have been performed within a frame or among frames, the designed algorithm utilizes inter-layer parallelism based on a group of macroblocks (GOM). Also, improved buffer management can be achieved by the proposed access unit (AU) based decoding architecture for enabling GOM-based inter-layer multithreading architecture. The proposed multithreading architecture has three properties: (1) scalable to the number of SVC layers, (2) no additional coding delay, and (3) no additional memory requirement. Experimental results show that the proposed multithreading architecture speeds up the decoding time of 3-layer extended spatial scalability sequences by about 36% on average, 3-layer coarse grain scalability 50%, and 5-layer medium grain scalability by about 102%, respectively, compared to a single-threaded SVC decoder.

Keywords: Scalable Video Coding (SVC), Multithreading algorithm, Access unit based decoding.

1 Introduction

Recently video communication services, such as Internet Protocol television (IPTV), mobile IPTV, mobile broadcasting, and multi-screen media service, through various networks and devices have gained growing global interests [1]-[5]. But all these services need to guarantee the minimum quality of service. To meet these service requirements, the new standard which can provide the best quality of service at any environment are required not only in channel coding area but also in source coding area.

To satisfy the increasing industrial needs, ISO/IEC JTC1 SC29 WG11 MPEG and ITU-T SG16 Q.6 VCEG has introduced the new international standard named Scalable Video Coding (SVC) on the end of 2007 as an amendment 3 of MPEG-4

AVC/H.264 [6]. The SVC can provide a scalable bitstream which supports multiple sub-bitstreams under various spatial, temporal, and quality resolutions [6], [7]. The previous international video coding standards like MPEG-2/H.262 [8], H.263+ [9], and MPEG-4 Visual [10] have already supported various scalable tools. But the previous scalable functionalities of those standards have rarely been used, unfortunately, in the commercial market because of two major reasons [7], [11]. First, the scalable techniques comes along with a significant loss in coding efficiency and a large increase in decoder complexity, compared to those existing alternative solutions. Second, the characteristics of traditional video transmission system were not adequate enough for scalable service.

To overcome drawbacks of the previous scalable tools, the MPEG-4 AVC/H.264 SVC has increased coding efficiency by exploiting inter-layer intra, residual, and motion correlation, also has decreased decoder complexity by introducing single-loop decoding [6], [7]. The complexity of SVC is still high compared to single-layer MPEG-4 AVC/H.264 coding, but additional complexity of SVC mainly arises from considerable memory access for loading and storing data of reference layer (RL) and performing up-sampling operation of inter-layer intra, residual, and motion prediction. It means, if memory-efficient buffer management and complexity reduction on up-sampling process can be achieved, the major limitation of MPEG-4 AVC/H.264 SVC can be solved and thus SVC can be used widely. Especially, fast and memory-efficient up-sampling techniques are essential for real-time Scalable High profile decoder and for low-power SVC decoder on mobile environment. In order to speed up SVC up-sampling operation, Kim, et al. proposed fast and memory-efficient up-sampling methods for extended spatial scalability (ESS), where intra and residual up-sampling operations for ESS are selectively performed by using macroblock (MB) information of enhancement layer (EL) if necessary [12]. Yi, et al. proposed an access unit (AU) based SVC decoding architecture and common residual buffer structure for inter-layer residual prediction (ILRP), in order to significantly reduce memory access and consumption for residual prediction of spatial scalability (SS), coarse grain scalability (CGS), and medium grain scalability (MGS) with and without transform coefficient level prediction [13]. Chunag, et al. analyzed bandwidth overhead of typical SVC decoding and proposed a MB-based up-sampling for spatial scalability and a layer-interleaving decoding scheme for quality scalability in the hardware platform [14]. On the other hand, multi-threaded decoding algorithms for single-layer MPEG-4 AVC/H.264 decoder and various optimization techniques for designing multi-threaded applications have been studied widely [15-21]. The wavefront algorithm can decode several independent MBs concurrently by rearranging the data partition and task scheduling [15]. Chong, et al. presented preparting technique coupled with run-time MB level scheduling [16]. These methods, however, need considerable number of synchronizations which hurt the overall performance gain obtained from parallel processing [17]. To reduce synchronization overhead, Nishihara et al. proposed a task-parallel approach where a coarse, flexible partitioning adapted for the MPEG-4 AVC/H.264 decoding functions, such as motion compensation, deblocking filtering (DF), and variable length decoding, was developed [18], [19]. Su, et al. proposed a parallel algorithm for SVC decoder, which only supports multi-core stream processor and 3-layer spatial

scalability [21]. Until now, multithreading algorithms for SVC decoder with full function have been rarely studied.

In this paper, AU-based SVC decoding architecture is proposed to implement the proposed multithreading algorithm in the SVC decoder and to reduce memory access and consumption. And, the inter-layer multithreading architecture based on a group of macroblocks (GOM) for real-time Full HD MPEG-4 AVC/H.264 Scalable High profile decoder is presented.

This paper is organized as follows. In section 2 an AU-based SVC decoding architecture is presented. The proposed inter-layer multithreading architecture, and MB-based intra and residual up-sampling methods which are slightly modified version from [12] are presented in section 3. The proposed multithreading algorithm is compared to a single-threaded architecture, in terms of decoding speed, by experiments in section 4, and section 5 concludes the paper.

2 AU-Based SVC Decoding Architecture

An improved AU-based SVC decoding architecture is proposed for considerable reducing memory consumption and bandwidth of ILRP [13]. The proposed method can be applied to both single-threaded (ST) SVC decoder and multi-threaded (MT) one. Note that the proposed algorithm in this section becomes underlying architecture to form the proposed GOM-based inter-layer multithreading architecture.

The SVC supports three types of ILRP, such as transform coefficient level (TCL), transform coefficient (TC), and pixel residual (PR) prediction, for reducing residual signal among layers. Fig. 1 (a) shows one RL MB decoding flow from the viewpoint of ILRP in the JSVM that employs network abstraction layer unit (NALU)-based decoding [22].

In the Fig. 1, the Q^{-1} and IDCT mean inverse quantization and inverse integer discrete cosine transform, respectively. The proposed approach refers to the common residual buffer which stores one of TCL, TC and PR data instead of storing all residual data of RL for ILRP when decoding an EL as shown in the Fig. 1 (b). It is obvious

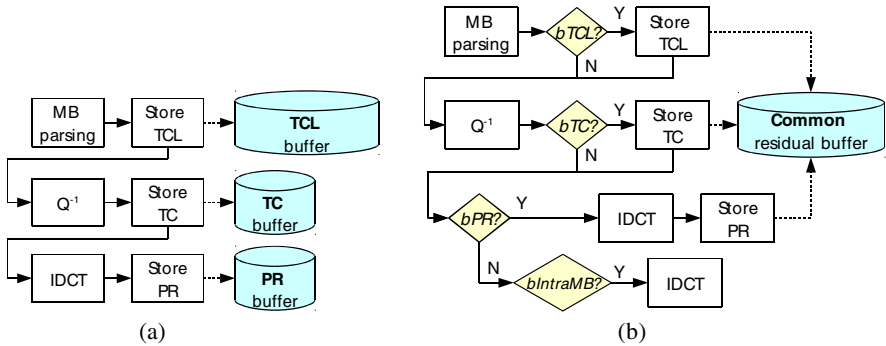


Fig. 1. (a) One RL MB decoding flow from the viewpoint of ILRP operation in the JSVM. (b) One RL MB decoding flow from the viewpoint of ILRP in the proposed AU-based decoding architecture.

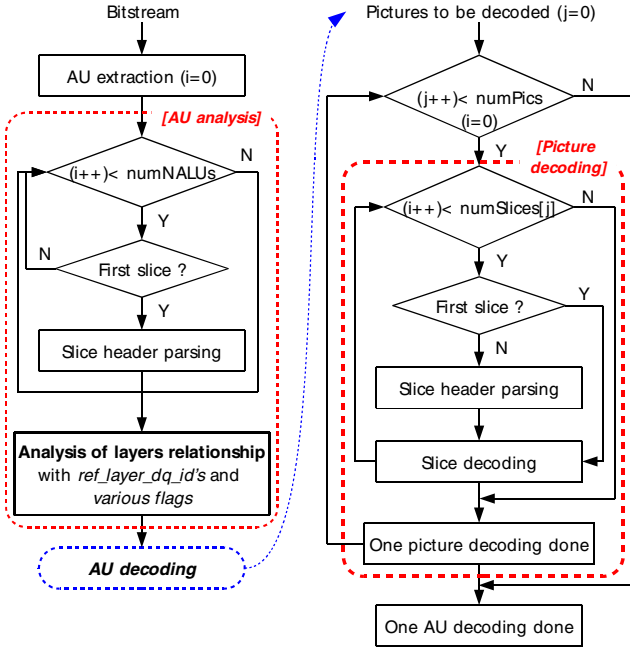


Fig. 2. AU-based SVC decoding architecture

that the proposed method can reduce unnecessarily considerable memory access and consumption in the course of RL decoding. Furthermore, AU-based decoding can eliminate unnecessary IDCT process to store PR data when decoding an RL.

In the Fig. 2, the flowchart of the newly developed AU-based SVC decoding architecture is shown where numNALUs and numPics represent the number of NALUs and pictures within one AU, respectively. The numSlices[j] means the number of slices within a j-th picture. AU-based decoding is performed by two steps: 1) analysis of layers relationship and 2) layers decoding. A SVC AU can be extracted by using the order decision process of NAL unit which is described in subclause 7.4.1.2 and G.7.4.1.2 of the SVC standard [6]. As shown in left part of Fig. 2, picture size, ref_layer_dq_id, and tcoeff_level_prediction_flag values of each layer can be translated by parsing sequence parameter sets and first slice header of each layer before decoding layers. By layer information analysis prior to layers decoding, dependencies among layers can be obtained and what RL data is necessary for ILRP of upper layers can be acquired in advance.

Fig. 3 shows the derivation process of three parameters for fast and memory-efficient ILRP before decoding an RL, where bTCL, bTC, and bPR represent flags whether inter-layer transform coefficient level, transform coefficient, and pixel residual prediction is used or not, respectively. Scalability type of EL can be identified by using picture size information of both EL and the corresponding RL.

As shown in Fig. 3, only one among three parameters should be equal to one. That means storing only one residual data out of three data is sufficient for an RL

decoding. Note that the bTCL, bTC, and bPR parameters can have different values among layers. In the analysis process, it is able to identify layers which are not referred by ELs by comparing all `ref_layer_dq_id` values and dependency-quality identifiers (DQIDs) of layers prior to AU decoding.

```

If EL is quality layer {
  If tcoeff_level_prefiction_flag of EL == 1 {
    bTCL = 1; bTC = bPR = 0; }
  else { // tcoeff_level_prediction_flag == 0
    bTC = 1; bTCL = bPR = 0; }
}
else { // EL is spatial layer
  bPR = 1; bTCL = bTC = 0;
}

```

Fig. 3. Three parameters derivation process for fast and memory-efficient ILRP before decoding a RL picture

As shown in right part of Fig. 2, layers decoding are performed with analyzed layers information. Note that the `numPics` represents the number of pictures to be decoded. That is, layer pictures which are not referred by ELs are not decoded. The three parameters derived for each layer are input into each layer decoder prior to picture decoding. Fig. 1 (b) shows one RL MB decoding flow from the viewpoint of ILRP in the proposed AU-based decoding architecture. When decoding an RL, only one among TCL, TC, and PR data is stored according to parameters bTCL, bTC, and bPR values. Also, if bPR is not equal to one and MB type is not intra, that is, `bIntraMB` is equal to zero, IDCT operation is not performed.

The proposed AU-based decoding architecture has four advantages: 1) considerable reduction of memory access and consumption by using only one common residual buffer and selective storing of residual data, 2) reducing computational complexity by effective identifying and skipping pictures not to be decoded, 3) reducing computational complexity by selective skipping unnecessary IDCT operation, and 4) providing easy framework for inter-layer multithreading architecture, which is shown in section 3. In the proposed algorithm, memory access and consumption for ILRP is reduced by half at least, compared to conventional method. That is because the maximum size of the common residual buffer is equal to a size of the existing TCL buffer.

3 GOM-Based Inter-layer Multithreading Architecture for SVC

Based on the algorithms of section 2, the fast and memory-efficient multithreading architecture for real-time Scalable High profile SVC decoder is proposed.

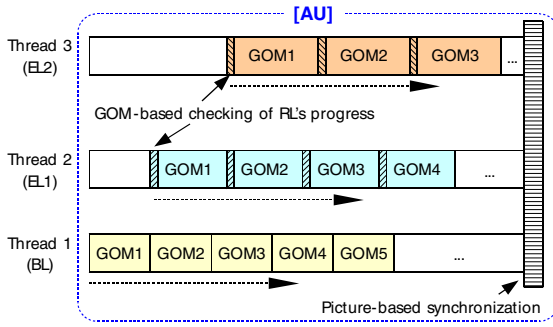


Fig. 4. GOM-based inter-layer multithreading architecture

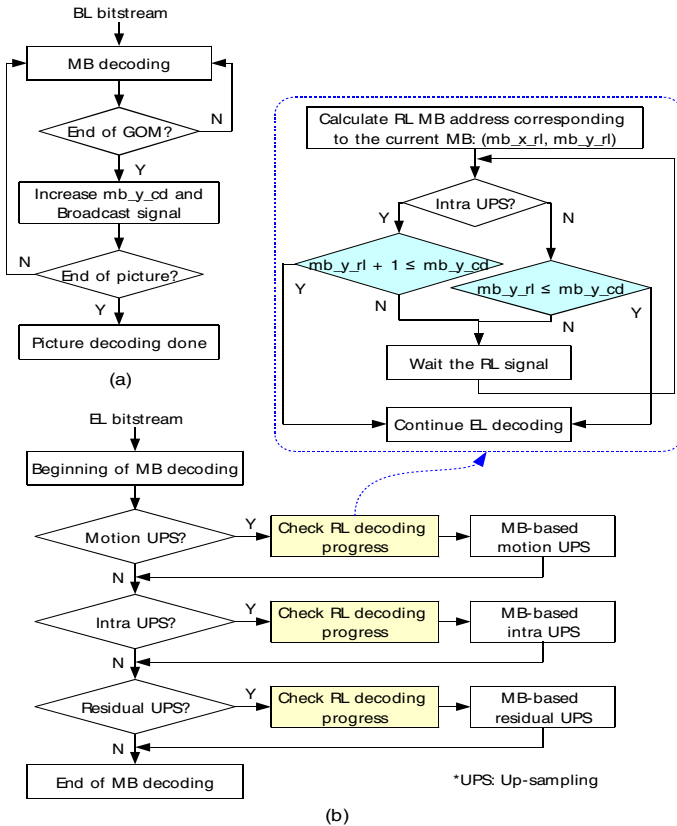


Fig. 5. GOM-based checking of RL decoding progress: (a) BL GOM decoding architecture, (b) EL MB decoding architecture including checking of RL decoding progress

Basically the proposed architecture assigns a thread to one layer decoding. That is, the picture decoding in the right part of Fig. 2 is performed by the corresponding layer threads. An example of the proposed architecture is shown in the Fig. 4, where BL, EL1, and EL2 represent base layer, enhancement layer 1, and enhancement layer 2, respectively. An EL MB can be decoded after decoding the corresponding RL MB since the SVC supports MB-based adaptive inter-layer prediction. Therefore, MB-based synchronization between RL and EL is required for multi-threaded layer decoding, which results in considerable synchronization overhead.

To significantly reduce synchronization overhead, the GOM-based checking of RL's decoding progress as shown in Fig. 4 and Fig. 5 is employed, where the GOM size can be different among layers. In this study, the GOM size of a layer set equal to the number of MBs in a MB row of the layer and thus the GOM size of ELs has the same value as one of BL in the quality scalability. On the contrary, the GOM size of each layer is different in the case of spatial scalability.

Decoding steps are as follows:

- (1) BL and RL thread increase mb_y_cd and broadcast completion signal whenever it completes GOM decoding [Fig. 5 (a)].
 - The mb_y_cd represents the mb_y address currently decoded in the RL.
 - Note that the signal should be broadcasted since one RL can be referenced by one or more EL.
- (2) An EL thread checks whether the RL GOM corresponding to the current MB was decoded or not before decoding a GOM of EL [Fig. 5 (b)].
 - The checking is performed only if the current MB uses inter-layer prediction.
 - If the corresponding GOM of RL is not yet decoded, the EL thread waits the GOM decoding completion signal broadcasted by the RL thread.
 - If the corresponding GOM of RL was decoded already, the EL thread continues GOM decoding.

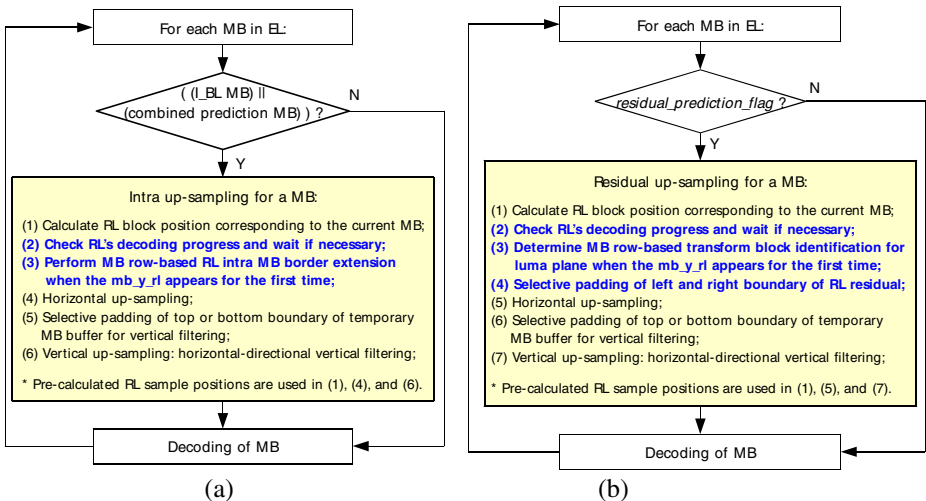


Fig. 6. (a) MB-based intra up-sampling method, (b) MB-based residual up-sampling method

The EL thread waiting overhead is not so high since RL decoding speed is faster than EL decoding speed in most cases. Also, if the EL thread should wait the completion signal in some cases, the waiting is occurred only in the first MB within a GOM of EL. To realize the proposed multithreading architecture, the existing MB-based intra and residual up-sampling methods are modified, which have hybrid architecture combining picture-based pre-processing and MB-based up-sampling [12].

Fig. 6 shows modified MB-based intra and residual up-sampling methods, respectively. In the Fig. 6, the combined prediction MB is a inter MB with two properties that `base_mode_flag` is equal to 1 and some region of RL block corresponding to the current MB contains intra pixels. The RL decoding progress checking and MB-based up-samplings of Fig. 5 (b) are combined to form complete module in the Fig. 6 (a) and Fig. 6 (b), respectively.

The picture-based pre-processing by EL used in the existing up-sampling methods is not possible in the proposed multithreading architecture since the RL picture is not yet fully decoded prior to EL decoding. To overcome the above problem and to deal with data dependency problem of intra padding process, MB row-based intra MB border extension method is introduced. The proposed intra padding is performed when the `mb_y_rl` value appears for the first time. Before performing intra padding operation, it should be checked the decoding progress of RL picture since intra MB padding operation requires right and bottom MB types and pixel data. That is, the RL picture should be decoded to $(mb_y_rl + 1)$ -th MB row before padding mb_y_rl -th MB row. Otherwise the EL thread should wait $(mb_y_rl + 1)$ -th MB row decoding completion signal broadcasted by the RL thread as shown in the Fig. 6 (b). On the other hand, determining transform block identification requires that the RL picture should be decoded to mb_y_rl -th MB row before determining that of mb_y_rl -th MB row. Consequently, step (2) and (3) is added to the existing intra up-sampling method in the Fig. 6 (a) and add step (2), (3), and (4) to the existing residual up-sampling algorithm in the Fig. 6 (b).

The proposed multithreading architecture has five properties: 1) applicable to both SVC encoder and decoder, 2) inherently scalable to the number of SVC layers, 3) no additional coding delay, 4) no additional memory requirement, and 5) really much faster than single-threaded SVC architecture, which is proved in following section.

4 Simulation Results

This section covers test conditions and comparison of decoding speed between ST and the proposed MT SVC decoder.

4.1 Test Conditions

To verify the proposed algorithms, a lot of SD (704x576) and Full HD (1920x1080p) sequences are coded with SS, CGS, and MGS configurations by using JSVM encoder version 9.19.8 [22]. Table 1 shows encoding parameters for each test configuration. Experiment was performed four times in a row in the workstation¹. The purpose of the

¹ Two Zeon X5482@3.2GHz CPU (quad-core), 4 GB DDR2 RAM, and Windows Vista SP2.

first replication was to load the program code and the bitstream into the disk cache and at least partially into the L2 cache [23]. The median value of three other replications was reported. All the experiments are performed by only exchanging MT and ST functions of section 3 in our optimized SVC decoder. The number of threads is the same as the number of SVC layers in the MT decoder experiment.

Note that the algorithms of section 2 and MB-based motion up-sampling method were included in both ST and MT decoder used in the experiment.

Table 1. Encoding parameters of test sequences

Parameters	Value	
Profile	Scalable High	
Frame rate [Hz]	Full HD (25), SD (30)	
# frames	Full HD Bluesky (217), Other Full HDs (300), SD (200)	
Intra period	32	
# reference frame	1	
ME mode	Fast log-search (Search mode: 4)	
Search range	Full HD (96), SD (64)	
Loop filter	On (0)	
Tools	8x8 trans., Adaptive inter-layer intra/motion/residual prediction	
Configuration A (3 layers SS)	720x480p/1280x720p/1920x1080p Entropy: CAVLC GOP: 8 (Hierarchical-B)	: 3-layer ESS with wide range of quality, which is a typical scenario for multi-screen media service
Configuration B (3 layers Full HD CGS)	1920x1080p Entropy: CABAC GOP: 4 (Hierarchical-B)	: High quality CGS with CABAC and low QP, which represents a seamless Full HD movie streaming service scenario in the best-effort network
Configuration C (3 layers SD CGS)	704x576 Entropy: CAVLC GOP: 8 (Hierarchical-B) tcoeff_level_prediction_flag = 1	: Medium quality CGS with MPEG-4/H.264 rewriting capability, which represents a backward-compatible SD movie streaming service scenario
Configuration D (5 layers Full HD MGS)	1920x1080p Entropy: CAVLC GOP: 4 (Hierarchical-B) Scan_idx:[0,15]/[0,2]/[3,6]/[7,10]/[11,15]	: 5-layer MGS with wide range of quality, which represents more accurate quality control scenario over CGS

4.2 Decoding Speed

Table 2 shows the decoding speed comparison between ST and the proposed MT SVC decoder with test configuration A, where fps and kbps represent frame per second and kilobits per second, respectively. In this scenario, MT decoder is faster than ST decoder by about 36.5% on average. The proposed MT decoder shows real-time decoding for all sequences and all QPs except one high quality (QP=18) Tractor sequence which has high texture background and camera motion. In fact, the high quality Tractor sequence is not adequate for consumer video service since the bitrate is too high, i.e., 46 Mb/s, as shown in the Table 2.

Table 3 and 4 show the comparison results with test configuration B and C, respectively. The MT decoder is faster than ST decoder by about 50.4% and 72.1% on average, respectively, in these scenarios. The decoding speed of test configuration B is relatively slower than the other configurations due to CABAC entropy decoding.

Table 2. The decoding speed comparison between a ST and the proposed MT decoder for test configuration A

Parameters	QP	PSNR[dB]	Bitrate[kbps]	ST[fps]	MT[fps]	Gain[%]
Bluesky (BS)	18	44.00	39352.41	19.03	25.28	32.84
	23	41.63	16643.46	27.00	36.02	33.41
	28	39.75	7996.60	34.20	46.49	35.94
	33	37.95	4773.65	36.62	50.20	37.08
Pedestrian (PE)	18	44.43	26020.70	21.95	29.42	34.03
	23	42.96	10548.51	28.99	40.22	38.74
	28	41.62	4670.44	34.04	48.01	41.04
	33	40.22	3078.02	36.62	52.22	42.60
Rushhour (RH)	18	43.61	25763.03	19.50	25.83	32.46
	23	43.02	9451.03	26.98	36.93	36.88
	28	42.00	3769.93	33.78	46.78	38.48
	33	41.02	2561.92	37.20	51.74	39.09
Sunflower (SF)	18	44.78	18620.32	25.95	34.57	33.22
	23	43.49	6903.74	35.86	49.39	37.73
	28	42.03	3111.94	41.73	57.79	38.49
	33	40.31	1944.26	43.93	61.32	39.59
Station (ST)	18	43.43	25718.22	24.22	30.67	26.63
	23	41.95	9061.29	37.80	52.28	38.31
	28	40.85	3156.49	43.12	61.67	43.02
	33	39.34	1826.15	45.45	64.71	42.38
Tractor (TR)	18	43.32	46305.29	15.39	20.11	30.67
	23	41.04	20056.58	21.22	28.07	32.28
	28	39.30	9456.63	26.53	35.80	34.94
	33	37.89	5877.10	29.12	39.36	35.16
Average	25.5	41.66	12777.82	31.09	42.70	36.46

Table 3. The decoding speed comparison between a ST and the proposed MT decoder for test configuration B

Seq.	QP	PSNR[dB]	Bitrate[kbps]	ST[fps]	MT[fps]	Gain[%]
BS	29/24/20	43.49	25790.61	19.70	29.89	51.73
PE	28/24/18	44.84	28386.58	18.95	27.82	46.81
RH	28/23/18	44.78	28113.36	18.28	26.53	45.13
TR	28/22/18	43.45	52549.38	12.82	20.22	57.72
Avg.	28/23/19	44.14	33709.98	17.44	26.12	50.35

Table 5 shows the comparison results for test configuration D. The proposed MT decoder is faster than ST decoder by about 101.7 % on average. In this MGS scenario, the ST decoder cannot decode all bitstreams in real-time. The MT decoder, however, shows real-time decoding for all sequences and all QPs.

Speed-up gains for four test configurations is quite different due to variable thread synchronization overhead and load imbalance between layers. Especially, the proposed multithreading architecture shows higher speed-up gain for CGS and MGS decoding since inherently quality scalability has more balanced load between layers than spatial scalability.

Table 4. The decoding speed comparison between a ST and the proposed MT decoder for test configuration C

Seq.	QP	PSNR[dB]	Bitrate[kbps]	ST[fps]	MT[fps]	Gain[%]
City	33/28/23	40.70	5636.54	123.16	207.05	68.11
Harbour	33/28/23	40.62	9740.48	93.15	144.85	55.50
Ice	33/28/23	44.73	2552.87	161.26	312.72	93.92
Soccer	33/28/23	41.55	5693.56	121.95	208.13	70.67
Avg.	33/28/23	44.14	5905.86	124.88	218.19	72.05

Table 5. The decoding speed comparison between a ST and the proposed MT decoder for test configuration D

Seq.	QP	PSNR[dB]	Bitrate[kbps]	ST[fps]	MT[fps]	Gain[%]
BS	24/18	44.79	44891.22	15.73	33.60	113.60
	29/23	42.24	20704.44	19.49	38.85	99.33
	34/28	40.13	10984.17	22.45	43.73	94.79
PE	24/18	45.07	34133.67	16.91	36.06	113.25
	29/23	43.24	12905.67	21.59	42.90	98.70
	34/28	41.76	6591.23	24.02	46.50	93.59
RH	24/18	44.82	35398.54	16.29	34.26	110.31
	29/23	43.33	11756.15	21.45	41.90	95.34
	34/28	42.17	5471.36	24.49	46.60	90.28
TR	24/18	44.14	54517.41	14.09	30.88	119.16
	29/23	41.64	24451.88	17.94	36.08	101.11
	34/28	39.69	12414.59	21.01	40.09	90.81
Avg.	29/23	42.75	22851.69	19.62	39.29	101.69

5 Conclusion

In this paper the proposed fast and memory-efficient multithreading architecture and AU-based SVC decoding architecture was described. The proposed AU-based SVC decoding architecture reduced memory access and consumption by half at least by analyzing relationship of layers. The proposed inter-layer multithreading architecture speeds up the SVC decoder up to 101% without additional coding delay and memory consumption, compared to a single-threaded decoder.

Although the implementation of the proposed multithreading architecture was done only in the SVC decoder with SS, CGS, and MGS, the proposed architecture can also be applied to SVC encoder. Furthermore, the multithreading architecture can be implemented on any general-purpose processor, such as x86 and ARM.

References

1. Wiegand, T., Noblet, L., Rovati, F.: Scalable Video Coding for IPTV services. *IEEE Trans. Broadcasting* 55(2), 527–538 (2009)
2. Schierl, T., Stockhammer, T., Wiegand, T.: Mobile video transmission using Scalable Video Coding. *IEEE Trans. Circuits Syst. Video Technol.* 17(9), 1204–1217 (2007)

3. ATSC Mobile DTV Standard, Part 7 - AVC and SVC Video System Characteristics: A/153 Part7:2009, ATSC (2009)
4. Choi, H., Shin, I.H., Lim, J.-S., Hong, J.W.: SVC application in advanced T-DMB. *IEEE Trans. Broadcasting* 55(1), 51–61 (2009)
5. Tan, P., Slevinsky, J.: Multi-screen IPTV enabling technologies and challenges. In: *Proc. IEEE Int. Conf. Consumer Electronics*, pp. 1–2 (2011)
6. Advanced video coding for generic audiovisual services: ITU-T Rec. H.264 and ISO/IEC 14496-10, ITU-T and ISO/IEC JTC 1 (2010)
7. Schwarz, H., Marpe, D., Wiegand, T.: Overview of the Scalable Video Coding extension of the H.264/AVC standard. *IEEE Trans. Circuits Syst. Video Technol.* 17(9), 1103–1120 (2007)
8. Generic Coding of Moving Pictures and Associated Audio Information Video: ITU-T Rec. H.262 and ISO/IEC 13818-2, ITU-T and ISO/IEC JTC 1 (2000)
9. Video Coding for Low Bit Rate Communication: ITU-T Rec. H.263, ITU-T (2005)
10. Coding of Audio-Visual Objects - Part 2 Visual: ISO/IEC 14496-2, ISO/IEC JTC 1 (2004)
11. Liu, H., Wang, Y.-K., Li, H.: A comparison between SVC and transcoding. *IEEE Trans. Consumer Electronics* 54(3), 1439–1446 (2008)
12. Kim, Y.-H., Yi, J.-Y., Choi, B.: Fast and memory-efficient up-sampling methods for H.264/AVC SVC with extended spatial scalability. *IEEE Trans. Consumer Electronics* 56(2) (2010)
13. Yi, J.-Y., Kim, Y.-H., Choi, B.: Fast and memory-efficient AU-based decoding method for H.264/AVC SVC. In: *Proc. CEWIT, Incheon, Korea* (2010)
14. Chuang, T.-D., Tsung, P.-K., Lin, P.-C., Chang, L.-M., Ma, T.-C., Chen, Y.-H., Chen, L.-G.: Low bandwidth decoder framework for H.264/AVC Scalable Extension. In: *Proc. IEEE Int. Symp. Circuits and Systems*, pp. 2960–2963 (2010)
15. van der Tol, E.B., Jaspers, E.G.T., Gelderblom, R.H.: Mapping of H.264 decoding on a multiprocessor architecture. In: *Proc. SPIE Conf. on Image and Video Communications and Processing*, pp. 707–718 (2003)
16. Chong, J., Satish, N., Catanzaro, B., Ravindran, K., Keutzer, K.: Efficient parallelization of H.264 decoding with macro block level scheduling. In: *Proc. IEEE ICME*, pp. 1874–1877 (2007)
17. Yang, S.-S., Wang, S.-W., Wu, J.-L.: A parallel algorithm for H.264/AVC deblocking filter based on limited error propagation effect. In: *Proc. IEEE ICME*, pp. 1858–1861 (2007)
18. Nishihara, K., Hatabu, A., Moriyoshi, T.: Parallelization of H.264 video decoder for embedded multicore processor. In: *Proc. IEEE ICME*, pp. 329–332 (2008)
19. Sihm, K.-H., Baik, H., Kim, J.-T., Bae, S., Song, H.J.: Novel approaches to parallel H.264 decoder on symmetric multicore systems. In: *Proc. IEEE ICASSP*, pp. 2017–2020 (2009)
20. Kim, D., Lee, V.W., Chen, Y.-K.: Image processing on multicore x86 architecture. *IEEE Signal Processing Magazine* 27(2), 97–107 (2010)
21. Su, Y.-C., Tsai, S.-F., Chuang, T.-D., Tsao, Y.-M., Chen, L.-G.: Mapping Scalable Video Coding decoder on multi-core stream processors. In: *Proc. Picture Coding Symposium*, pp. 1–4 (2009)
22. Reichel, J., Schwarz, H., Wien, M.: Joint Scalable Video Model 11 (JSVM 11): JVT, Geneva, CH, Doc. JVT-X202 (2007)
23. Lappalainen, V., Hallapuro, A., Hamalainen, T.D.: Complexity of optimized H.26L video decoder implementation. *IEEE Trans. CSVT* 13(7), 717–725 (2003)

Fast Mode Decision Algorithm for Depth Coding in 3D Video Systems Using H.264/AVC

Da-Hyun Yoon and Yo-Sung Ho

Gwangju Institute of Science and Technology (GIST)
261 Cheomdan-gwagiro, Buk-gu, Gwangju 500-712, Korea
{yoon,hoyo}@gist.ac.kr

Abstract. Complexity of multiview coding is proportional to the number of cameras. It makes difficult to implement multiview sequences in real applications. Thus, we propose a fast mode decision algorithm for both intra and inter prediction to reduce the computational complexity of H.264/AVC for depth video coding. By analyzing the depth variation, we classify the depth video into depth-continuity and depth-discontinuity regions. We determine a threshold value for classifying these regions by experiments. Since the depth-continuity region has an imbalance in the mode distribution, we limit the mode candidates. Experimental results show that our proposed algorithm reduces the encoding time by up to 78% and 84% for the intra and inter frames, respectively, with negligible PSNR loss and slight bit-rate increase, compared to JMVC 8.3.

Keywords: depth video coding, macroblock mode decision, depth compression.

1 Introduction

The development of 3DTV has realized the human dream of viewing scenes as if in the real world owing to advances in three-dimensional (3D) display technologies. Via 3DTV, the interactive selection of viewpoint and direction becomes possible within a certain operational range; this is referred to as free viewpoint TV (FTV). FTV has been widely used because it transmits and records all spatiotemporal information from the real world [1]. Multiview plus depth (MVD) is another framework used to represent 3D scenes; MVD has been used to synthesize intermediate views from captured images and depth maps. In a recent Moving Picture Experts Group (MPEG) meeting, MVD has received increased attention and has been discussed as a next-generation FTV format [2]. However, since the amount of data and complexity of MVD is proportional to the number of cameras, an efficient encoding method for 3D video must be developed. In order to utilize limited bandwidth and storage capacity efficiently, depth map sequences also have to be compressed in the same manner as texture videos.

In order to select the optimal coding mode, we use a rate-distortion optimization (RDO) technique [3] in JMVC, where J is the RD cost of the current mode, D denotes the distortion between the original and reconstructed macroblock (MB), R stands for the total bits of the MB header, motion vectors, and DCT coefficients, and λ is the Lagrange multiplier. The rate-distortion (RD) cost is calculated as follow

$$J=D+\lambda \cdot R \quad (1)$$

JMVC 8.3[4] is then used to select the best mode among 14 different macroblock modes: SKIP, Direct, Inter16×16, Inter16×8, Inter 8×16, Inter 8×8, Inter 8×8 Frext, Inter 8×4, Inter 4×8, Inter 4×4, Intra16×16, Intra 8×8, Intra4×4, and Intra PCM. The full search algorithm uses all modes to determine the optimal macroblock mode in terms of the RD cost, and the mode having the minimum RD cost is subsequently selected as the best mode. Unfortunately, the full search algorithm is time consuming, making it difficult to implement MVD in real applications. In this paper, we propose a fast mode decision scheme to reduce the complexity.

Depth sequences represent the distance between an object and a camera as a gray scale image. The image has a continuous area at an object and background and a discontinuous area at the boundaries between an object and background. Because the characteristics of depth sequences are very different from those of texture sequences, efficient coding algorithms specializing in depth sequences have been introduced. For example, Oh [5] re-uses motion information of the corresponding texture sequences in order to reduce the complexity in a motion estimation process and the bitrate for motion vector coding in the depth sequences. The fast mode decision method presented by Shin [6] uses a region analysis to reduce the complexity. Then, based on the RD cost correlation between neighboring views and the RD cost of different texture segmentation regions, a pre-decision of the SKIP mode was introduced by Whu [7]. Lee [8] skipped some blocks in the depth image to reduce the depth coding bitrate according to introducing an inter-view correlation of the texture image.

Approaches for fast depth video coding can be classified into two groups. One group exploits the correlation between color and depth video, and shares common information to reduce redundancy. The other group focuses only on the unique properties of depth video. Most fast depth coding algorithms belong to the first group. However, the second group is generally more available to various 3DTV applications because their algorithms only use unique properties of the depth video itself and is independent of the experimental framework.

The proposed algorithm consists of two parts: (1) the SKIP and intra modes are searched in the depth-continuity regions, whereas all modes are searched in the depth-discontinuity regions in the same manner as in JMVC; and (2) vertical, horizontal, DC, and diagonal down-left modes are implemented in the Intra 4×4 prediction of the depth-continuity regions. A threshold value is used to determine whether the region is homogeneous or not, and it is calculated adaptively according to the quantization parameter (QP).

2 Proposed Algorithm

2.1 Edge Classification

Depth sequences have different properties from those of texture sequences. Since the Intra 16×16, Intra 4×4, and SKIP modes are frequently selected as the best modes, motion vector search and the mode decision process can be skipped in homogeneous

regions. Before applying our proposed algorithm, we first separate a macroblock into continuous and discontinuous regions. For this task, the degree of variation of the depth value in a macroblock is defined by

$$f(x,y) = \frac{1}{16^2} \sum_{i=1}^{16} \sum_{j=1}^{16} (r(i,j) - m_{x,y})^2 \quad (2)$$

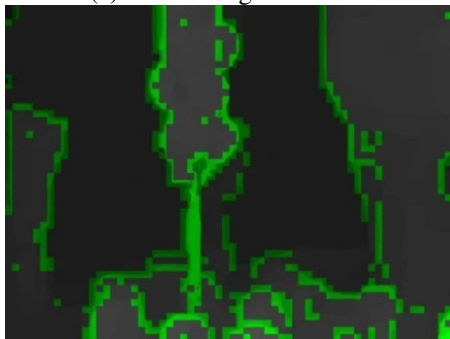
where coordinate (x,y) is the position of the current macroblock, $r(i,j)$ is the depth value at a relative coordinate (i,j) in the current macroblock, and $m_{x,y}$ is the mean value of the current macroblock. If the depth values drastically change, the value of $f(x,y)$ is large; thus, the value of $f(x,y)$ is large in discontinuity regions. Conversely, in continuity regions, where the depth values are almost fixed, the value of $f(x,y)$ is small. Therefore, we use a threshold value (T) to determine whether the current macroblock is located at a boundary or not. In Fig.1, macroblocks shown in green represent depth-discontinuity regions with $f(x,y) > T$. Compared to color sequence, depth-discontinuity that is filled with green macroblock regions are adequately detected. So, using variation is easy way to separate depth-continuity regions.



(a) color image of "Balloon"



(b) color image of "Book Arrival"



(c) depth image of "Balloon"



(d) depth image of "Book Arrival"

Fig. 1. Segmentation of discontinuity regions ($T = 30$)

2.2 Analysis of Mode Selection

The macroblock mode distribution of the depth-continuity regions is different from that in depth-discontinuity regions. Figure 1 compares the mode distribution between the depth-continuity and depth-discontinuity regions. In the depth-continuity regions, there is a severe imbalance in the mode distribution as most macroblocks are encoded by the SKIP and intra modes. However, in the depth-discontinuity regions the mode is balanced; therefore, we use this property to design a fast mode decision algorithm. Table 1 presents the encoding configurations for Fig. 2 and Fig. 3.

Table 1. Encoding configurations

Parameter	Setting
<i>Sequence</i>	Balloon
<i>Threshold</i>	30
<i>QP</i>	32
<i>View</i>	1
<i>Time</i>	0

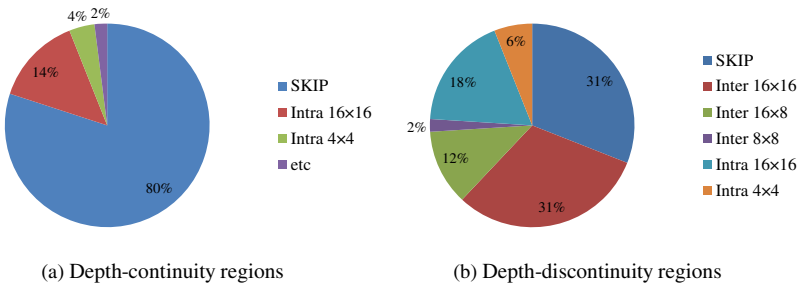


Fig. 2. Macroblock mode distribution

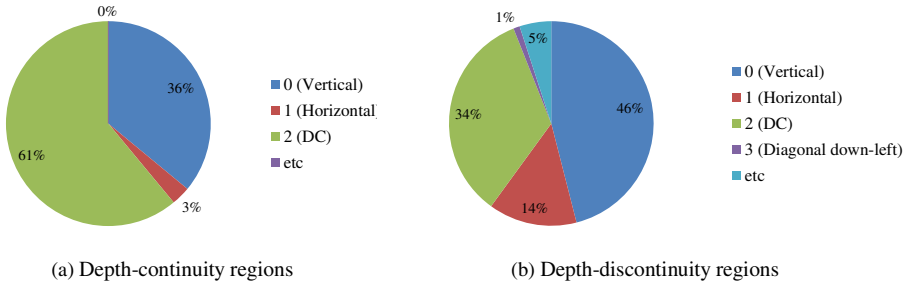


Fig. 3. Intra prediction mode distributions

Since mode 0, mode 1, mode 2, and mode 3 comprise close to 100% of the depth-continuity regions, as shown in Fig. 3, we consider these modes among the nine prediction modes selected as candidate modes set in the Intra 4x4 prediction. However, since the mode distributions for all four modes in the Intra 16x16 prediction are similar,

we do not need to consider the candidate mode set in the Intra 16×16 prediction. Therefore, we calculate the SKIP mode, the candidate mode set in the Intra 4×4 prediction, and all four modes in the Intra 16×16 prediction in order to determine the best mode in the depth-continuity regions. Since the distribution for all modes varies in the depth-discontinuity regions, we use a conventional mode decision method in JMVC.

2.3 Fast Depth Coding Algorithm

The algorithm developed by Peng [9] empirically indicates that the threshold value that discriminates between the depth-continuity and depth-discontinuity should be set to 30. If $f(x,y)$ is less than or equal to the threshold value, a macroblock is determined to be a continuous region. Since the SKIP and intra modes are used as candidate modes in continuous regions, we can reduce the complexity by skipping the inter mode in the mode decision. If $f(x,y)$ is larger than a threshold value, a macroblock is determined to be a discontinuous region, and all macroblock mode decisions including the inter mode decision are implemented. However, in low QPs, Peng's algorithm does not keep detailed boundary regions because it uses a fixed threshold value. To solve this problem, we use adaptive threshold values for the intra and inter predictions.

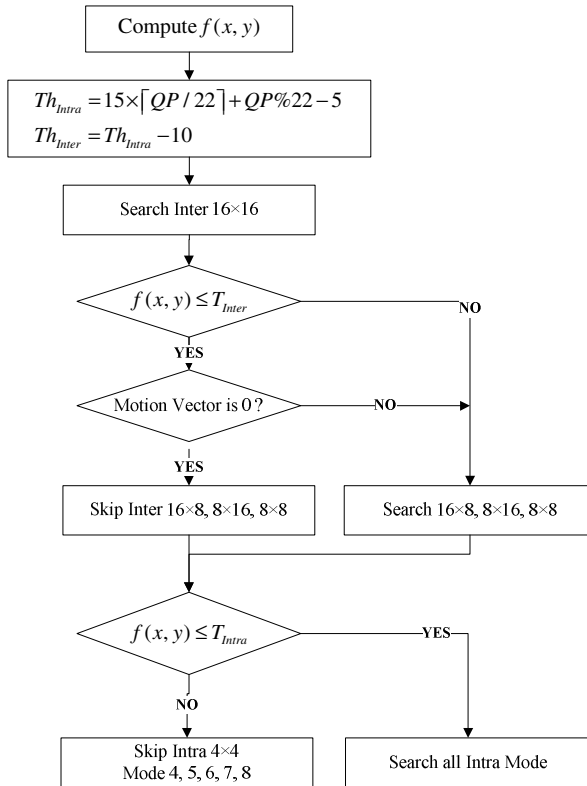


Fig. 4. Flowchart of the proposed algorithm

Figure 4 presents a flowchart of the proposed algorithm. After $f(x,y)$ is calculated at each macroblock, the threshold values for the intra and inter predictions are calculated according to the QP. In order to guarantee the coding efficiency at all QP ranges, we assign small and large threshold values in the low and high QP values, respectively. Since the SKIP of the inter mode decision induces additional bits from the wrong determination, the threshold value for the inter mode (T_{Inter}) is lower than that for the intra mode (T_{Intra}). After SKIP and inter 16×16 are performed, If $f(x,y) < T_{Inter}$ and the motion vector of inter 16×16 equals to 0, calculating inter 16×8, inter 8×16, and inter 8×8 to find the best mode of the current macroblock is skipped. Otherwise, we calculate inter 16×8, inter 8×16, inter 8×8. Next, if $f(x,y) < T_{Intra}$, mode decisions for mode 0, mode 1, mode 2, and mode 3 in the Intra 4×4 are performed. Otherwise, all modes are calculated in order to find the best intra prediction mode.

3 Experimental Results

In order to evaluate the efficiency of the proposed algorithm, we performed experiments on several depth sequences having 1024×768 resolutions. All test sequences have 50 frames, and we implemented our proposed algorithm on reference software JMVC 8.3. The detailed encoding parameters for the reference software are summarized in Table 2.

Table 2. Experimental conditions

Parameter	Setting
<i>Reference software</i>	JMVC 8.3
<i>Profile</i>	FRExt
<i>Depth sequences</i>	Book Arrival, Love Bird, and Newspaper
<i>Resolution</i>	1024×768
<i>Number of encoded frames</i>	50
<i>Search range</i>	±32
<i>QP</i>	22, 27, 32, 37
<i>Symbol mode</i>	1 (CABAC)
<i>GOP size</i>	Intra frame : 1 Inter frame : 8
<i>Intra period</i>	Intra frame : 1 Inter frame : 16

Table 3 shows that there are bitrate increases within $\pm 0.04\%$ and PSNR reductions up to -0.13 dB. Compared to the full search algorithm, up to 78.73% of the encoding time is saved; sufficiently compensating for the significant bitrate increase and decreased PSNR. Because the “Love Bird” depth sequence has many continuous background regions, it shows better performance compared to the other sequences.

Table 4 and Table 5 compare Peng’s algorithm and the proposed algorithm. There is a negligible bitrate increment and PSNR reduction of up to 1.57% and 0.26 dB in the proposed algorithm. The average PSNR difference is decreased by 0.12 dB and

Table 3. Comparison of intra coding performance in terms of Δ PSNR, Δ BR, and Δ TS

Sequences	QP	Δ PSNR (dB)	Δ BR (%)	Δ TS (%)	BDBR (%)	BDPSNR (dB)
Book Arrival	22	-0.11	-0.04	-8.41	0.00	-0.01
	27	-0.13	0.00	-11.06		
	32	-0.04	0.00	-11.47		
	37	0.01	0.00	-14.28		
Love Bird	22	-0.05	0.00	-78.73	-0.05	0.003
	27	-0.06	0.00	-70.03		
	32	-0.05	0.00	-74.46		
	37	-0.00	0.00	-76.43		
Newspaper	22	-0.03	-0.01	-27.43	0.12	-0.007
	27	-0.09	-0.01	-32.50		
	32	0.03	0.00	-29.54		
	37	-0.01	0.00	-29.58		
Average		-0.05	-0.01	-38.66	0.02	0.00

Table 4. Inter coding performance in terms of Δ PSNR, Δ BR, and Δ TS for Peng's algorithm

Sequences	QP	Δ PSNR	Δ BR	Δ TS	BDBR	BDPSNR
Book Arrival	22	-0.32	-1.42	-48.40	5.23	-0.174
	27	-0.41	-3.96	-52.64		
	32	-0.30	-3.98	-52.42		
	37	-0.26	-0.85	-40.37		
Love Bird	22	-0.28	-2.08	-92.52	2.5	-0.092
	27	-0.20	-1.39	-85.49		
	32	-0.16	-2.21	-85.32		
	37	-0.14	-0.65	-85.80		
Dog	22	-0.76	-6.03	-87.52	8.54	-0.298
	27	-0.80	-10.81	-86.68		
	32	-0.51	-9.28	-85.76		
	37	-0.29	-2.40	-84.44		
Kendo	22	-0.61	2.25	-83.99	17.75	-0.95
	27	-0.72	2.02	-84.45		
	32	-0.76	3.73	-84.75		
	37	-1.29	1.26	-84.14		
Pantomime	22	-0.59	5.80	-94.10	42.31	-1.665
	27	-1.09	8.21	-94.22		
	32	-0.94	22.04	-94.23		
	37	-3.11	4.23	-94.19		
Average		-0.22	-0.68	-83.79	15.27	-0.65

0.22 dB in the proposed algorithm and Peng's algorithm, respectively; the bitrates are reduced by an average of 0.37% and 0.68% in the two algorithms. In order to measure the overall improvement of the proposed method, the Bjontegaard delta bitrate (BDBR) and PSNR (BDPSNR) were used. BDBR and BDPSNR mean bitrate and PSNR under equivalent PSNR and bitrate, respectively.

Table 5. Inter coding performance for the proposed algorithm

Sequences	QP	Δ PSNR	Δ BR	Δ TS	BDBR	BDPSNR
Book Arrival	22	-0.22	1.57	-76.30	4.08	-0.156
	27	-0.20	0.51	-79.01		
	32	-0.12	-0.65	-68.95		
	37	-0.09	-0.77	-66.61		
Love Bird	22	-0.13	0.40	-81.57	0.53	-0.023
	27	-0.05	-0.79	-63.95		
	32	-0.01	-0.64	-62.07		
	37	-0.01	0.09	-66.33		
Dog	22	-0.17	-0.13	-84.0	2.26	-0.083
	27	-0.12	-0.49	-83.35		
	32	-0.05	0.10	-81.35		
	37	-0.04	-0.17	-82.65		
Kendo	22	-0.26	0.55	-79.77	2.6	-0.174
	27	-0.19	0.16	-78.90		
	32	-0.14	-0.80	-76.72		
	37	-0.07	-0.81	-74.89		
Pantomime	22	-0.19	-1.18	-79.61	1.78	-0.062
	27	-0.14	-2.19	-78.66		
	32	-0.09	-1.41	-76.68		
	37	-0.03	-0.81	-74.38		
Average		-0.12	-0.37	-75.79	2.25	-0.10

$$\Delta PSNR = PSNR_{proposed} - PSNR_{original} \quad (3)$$

$$\Delta TS = \frac{Time_{proposed} - Time_{original}}{Time_{original}} \times 100 \quad (4)$$

$$\Delta BR = \frac{Bitrate_{proposed} - Bitrate_{original}}{Bitrate_{original}} \times 100 \quad (5)$$

In terms of BDBR and BDPSNR, the proposed algorithm generates low bitrate and PSNR degradation compared to Peng's algorithm, which used a fixed threshold value of 30. Since low QP sequences are more sophisticated than high QP sequences, the inaccurate prediction process induced from skipping modes caused the quality degradation. Because Peng's algorithm skips many modes regardless of the macroblock properties, a bitrate increase and PSNR loss occur. In terms of BDBR and BDPSNR, Peng's algorithm represents BDBR increase and BDPSNR loss up to 42% and -1.6

dB respectively. This results show that Peng's algorithm is inefficient. However, experimental results show that the proposed algorithm is improved about 2.25% and -0.10 dB in terms of BDBR and BDPSNR respectively. Thus, experimental results in terms of BDBR and BDPSNR mean that the proposed algorithm maintains the quality of depth sequence coding and bitrate value compared to the Peng's algorithm. The proposed method reduces the encoding time an average of 75.79%. Figure 5 shows that BDBR and BDPSNR values are sensitive to T_{inter} and that the smaller the T_{inter} the smaller the bit increase and PSNR degradation.

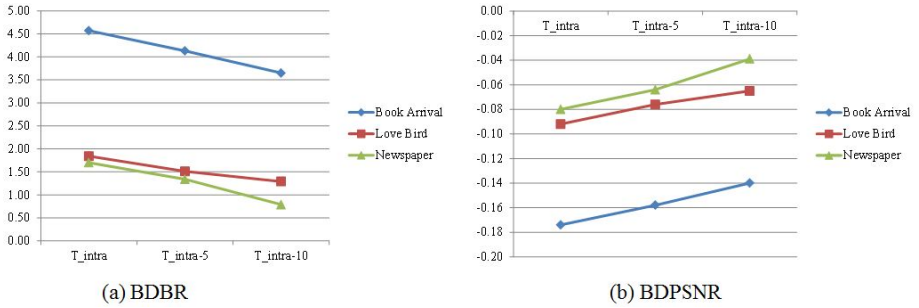


Fig. 5. BDBR and BDPSNR change according to T_{inter} value

To synthesize the intermediate virtual view, we used the 4th and 6th views in Newspaper and “Love Bird”, and the 7th and 9th views in “Book Arrival”

The results of PSNR differences between the synthesized view and original sequences compared to full search algorithm are shown in Table 7; VSRS rendering software was used for this experiment [10]. Both Peng's and the proposed algorithms have negligible differences in the range from -0.02 dB to 0.06 dB, indicating that both algorithms maintain rendering quality. Figure 6 shows that there is no significant rendering quality degradation between the full search and the proposed algorithm.



(a) Full search algorithm

(b) Proposed algorithm

Fig. 6. Synthesized results of “Book Arrival sequence”

Table 6. Δ PSNR of the synthesized result

	QP	22	27	32	37
Peng's algorithm	Book Arrival	0.01	0.01	0.06	-0.01
	Love Bird	0.01	0.00	0.00	-0.01
	Newspaper	0.00	0.00	0.00	0.00
Proposed algorithm	Book Arrival	0.02	0.02	0.06	-0.02
	Love Bird	0.01	0.01	-0.01	0.00
	Newspaper	0.00	-0.01	0.00	0.00

4 Conclusions

Although MVC coding is time consuming, it is still the most effective way to represent 3D scenes. The property of depth images is different from that of texture images. Since depth-discontinuity regions have imbalanced macroblock mode distributions, we propose a fast depth video coding algorithm using a threshold value that determines the depth-continuity or depth-discontinuity adaptive to the QP. If the variation is lower than the inter threshold value and motion vector is equal to zero, the proposed algorithm uses the SKIP and intra modes with no motion estimation or compensation; if the variation is smaller than the intra threshold value, the proposed algorithm performs the intra prediction using only mode 0, mode 1, mode 2, and mode 3. Our algorithm reduce the encoding time up to 78% and 84% for the intra and inter frames, respectively, with no significant degradation for the PSNR and rendering quality, or bitrate increment.

Acknowledgments. This research was supported by the MKE(Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency). (NIPA-2011-(C1090-1111-0003)).

References

1. Benzie, P., Watson, J., Surman, P., Rakkolainen, I., Hopf, K., Urey, H., Sainov, V., Kopylow, C.V.: A Survey of 3DTV displays: techniques and technologies. *IEEE Transactions on Circuits and Systems for Video Technology* 17(7), 1647–1658 (2007)
2. Mori, Y., Fukushima, N., Fujii, T., Tanimoto, M.: View generation with 3D warping using depth information form FTV. In: 3DTV Conference, pp. 229–232 (2008)
3. Sullivan, G.J., Wiegand, T.: Rate-distortion optimization for video compression. *IEEE Signal Process* 15, 74–90 (1998)
4. Joint multiview coding (JMVC) 8.3, <http://garcon.ient.rwth-aachen.de>
5. Oh, H., Ho, Y.-S.: H.264-Based Depth Map Sequence Coding Using Motion Information of Corresponding Texture Video. In: Chang, L.-W., Lie, W.-N. (eds.) *PSIVT 2006. LNCS*, vol. 4319, pp. 898–907. Springer, Heidelberg (2006)

6. Shin, K., Chun, S., Chung, K.: A fast mode prediction of multi-view video coding using region analysis. *Digital Content, Multimedia Technology and its Applications*, 87–90 (2010)
7. Whu, W., Jiang, W., Chen, Y.: A fast inter mode decision for multiview video coding. *Consumer Electronics*, 689–694 (2011)
8. Lee, J., Wey, H., Park, D.: A fast efficient multiview depth image coding method based on temporal and inter-view correlations of texture images. *IEEE Transactions on Circuits and Systems for Video Technology* 99, 1–4 (2011)
9. Peng, Z., Yu, M., Jiang, G., Si, Y., Chen, F.: Virtual view synthesis oriented fast depth video encoding algorithm. In: *International Conference on Industrial and Information Systems*, pp. 204–207 (2010)
10. ISO/IEC JTC1/SC29/WG11 M16090: View synthesis algorithm in view synthesis reference software 2.0 (VSRS 2.0) (2009)

Improved Diffusion Basis Functions Fitting and Metric Distance for Brain Axon Fiber Estimation

Ramón Aranda¹, Mariano Rivera¹, and Alonso Ramírez-Manzanares²

¹ Centro De Investigación en Matemáticas, Guanajuato, Gto, México, 36240

² Universidad de Guanajuato, Departamento de Matemáticas, Guanajuato, Gto, México, 36240

arac@cimat.mx, mrivera@cimat.mx, alram@cimat.mx

Abstract. We present a new regularization approach for Diffusion Basis Functions fitting to estimate *in vivo* brain the axonal orientation from Diffusion Weighted Magnetic Resonance Images. That method assumes that the observed Magnetic Resonance signal at each voxel is a linear combination of a given diffusion basis functions; the aim of the approach is the estimation of the coefficients of the linear combination. An issue with the Diffusion Basis Functions method is the overestimation on the number of tensors (associated with different axon fibers) within a voxel due to noise, namely, the over fitting of the noisy signal. Our proposal overcomes such an overestimation problem. In additionally, we propose a metric to compare the performance of multi-fiber estimation algorithms. The metric is based on the Earth Mover's Distance and allows us to compare in a single metric the orientation, size compartment and the number of axon bundles between two different estimations. The improvements of our two proposals is shown on synthetic and real experiments.

1 Introduction

Water diffusion estimation has extensively been used in recent years as an indirect way to infer axon fiber pathways and this, in turn, has made the estimation of fiber connectivity patterns *in vivo* one of the most challenging goals in neuroimaging. For this purpose, a special Magnetic Resonance Imaging (MRI) technique named Diffusion Weighted MRI (DW-MRI) is used. This imaging technique allows one to estimate the preferred orientation of the water diffusion in brains which, in the white matter case, is usually constrained along the axon orientations. This information is very useful in neuroscience research due to changes that occur in the neural connectivity patterns with neurological disorders and, in general, with brain development [1,2].

The water diffusion angular variation, within cerebral tissue, has been summarized in most medical applications by Diffusion Tensor Magnetic Resonance Images (DT-MRI) [3,4]. In the case of brain images, DT-MRI produces a tensor field that indicates the local orientation of nerve bundles. The local orientation of the corresponding nerve bundle is estimated from the orientation of

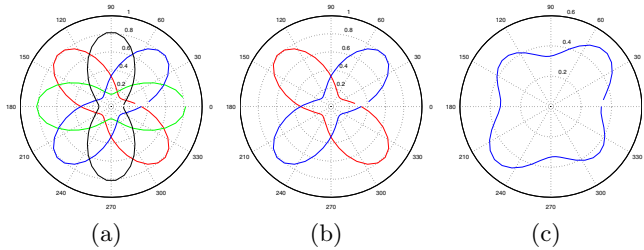


Fig. 1. 2D scheme of the diffusion basis functions, the narrow part of the signals correspond to the orientation of the diffusion peaks. (a) Set of 4 DBFs, (b) Subset of two DBFs, (c) Signal obtained by adding the two DBFs of (b).

the first eigenvector of a tensor, the one associated with its largest eigenvalue. This orientation is known as the Principal Diffusion Direction (PDD)[5]. In [6] Stejskal–Tanner presented a mono-exponential model of the decayed MR signal. Nowadays, it is well known that this technique is limited by the partial volume effect; i.e. diffusion at voxels with crossing fibers or bifurcations can not be represented by a single tensor.

Given the evident difficulty of DT-MRI for dealing with more than one diffusion direction per voxel, there has been proposed more sophisticated parametric models; see for instance refs. [7,8,9]. In particular, Multi-Tensor Diffusion MRI (MTD-MRI) can deal with fibers that split, merge or cross [7]. For instance, Tuch et.al [10] proposed to use the Gaussian Mixture Model (GMM). This model explains better the diffusion phenomenon for two or more fibers in a given voxel k , as:

$$S_{i,k} = S_{0,k} \sum_{j=1}^J \beta_{j,k} \exp(-bg_i^T T_{j,k} g_i), \quad (1)$$

where $S_{0,k}$ is the signal without diffusion, b is a constant acquisition parameter, $g_i = [g_{x_i}, g_{y_i}, g_{z_i}]^T$ is a unitary vector which indicates the direction on which is applied an independent magnetic gradient, with $i = 1, 2, \dots, M$ and M is the total number of applied gradients; $T_{j,k}$ is the j -th tensor, a 3×3 symmetric positive definite matrix; $S_{i,k}$ is the DW-MR signal measured when g_i was applied. β_j is the contribution of tensor $T_{j,k}$ and it is constrained by $\beta_j \in [0, 1]$ and $\sum_{j=1}^J \beta_j = 1$. Finally, J indicates the total number of tensors within the voxel. Thus, equation (1) shows the relationship between the signal without diffusion and the signal with diffusion on the direction g_i . Solving (1) is computationally expensive because it leads to a non linear optimization problem. Note that it requires to first estimate the number of axon bundles, J (*solve the model selection problem*), then to estimate all the others unknowns.

In [11] the authors proposed a strategy for solving the inverse problem stated in (1). They avoided the non-linear optimization problem by using a fixed set of Diffusion Basis Functions (DBF). The basis dictionary is, by definition, incomplete because it is computed on a discretization of the orientational 3D space. Then, the DW-MRI is approximated at each voxel by a linear combination of DBFs as it is shown in Figure 1:

$$S_{i,k} = \sum_{j=1}^N \alpha_j \phi_{i,j}, \quad (2)$$

with $\alpha_j \geq 0$, where the j -th DBF is defined by

$$\phi_{i,j} = S_{0,k} e^{(-bg_i^T \bar{T}_j g_i)}. \quad (3)$$

The coefficient $\phi_{i,j}$ can be understood as the diffusion weighted signal corresponding to the gradient g_i when a single perfect fiber is oriented with the largest engenvector of the base (fixed) tensor \bar{T}_j . Based on this approach, the GMM model (1) is transformed to a programming (optimization) problem. In particular, the non-negative least-squares (NNLS) formulation corresponds to [12,13]:

$$\begin{aligned} \min_{\alpha \in \mathbb{R}^N} \quad & \|\Phi\alpha - S\|_2^2 \\ \text{subject to} \quad & \alpha \geq 0 \end{aligned} \quad (4)$$

where $\Phi = \{\phi_{i,j}\}_{i=1,2,\dots,M,j=1,2,\dots,N}$ and $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_N]^T$ is the unknown vector of the linear system. It has been reported that this NNLS based DBF approach is prone to overestimate the number of fibers [14].

2 Sparse Diffusion Basis Functions

In this section, we propose a new regularization method for the DBF fitting to recover axonal intra-voxel information. As mentioned before, the solver in (4) is prone to recover more basis tensors than the actual number of tensors, that is to say, vector α is not so sparse. However, we note in our experiments that the solution given by solving (4) is close to the actual one. In the following, we say that α_j is active if $\alpha_j > 0$ and inactive if $\alpha_j = 0$, in other words, α_j is active if it contributes to the linear combination in (2). By experience, we know that the number of axon bundles within a voxel is small (1, 2 or 3) so that α must be sparse; i.e. the signal is approximated with a reduced number of component base signals. Thus, we propose to solve:

$$\begin{aligned} \min_{c \in \mathbb{R}, \alpha \in \mathbb{R}^N} \quad & \|c\Phi\alpha - S\|_2^2 + \lambda \|\alpha\|_2^2 \\ \text{subject to} \quad & \alpha \geq 0, \end{aligned} \quad (5)$$

using as starting point the solution given by (4); *i.e.*, the solution computed with $\lambda = 0$.

In our approach, sparse α vectors are promoted by the combined effect of: the second regularization term in (5), the non-negativity constraint and the scale factor c . The c parameter scales the solution, so that the error is reduced. The solution is computed by alternating the minimization with respect to (w.r.t.) α and c . In particular, the minimization w.r.t. α is achieved with a Gauss-Seidel scheme, where the non-negativity constraint is satisfied by projecting to zero

the negative α_j values in each iteration. This has the drawback that it requires several iterations to compute the optimum α . To accelerate the convergence, we iterate the Gauss-Seidel solver a few times alternated with a subspace minimization strategy [15]. On the other hand, the minimization w.r.t. c leads us to the close formula:

$$c = \frac{(\Phi\alpha)^T S}{\alpha^T \Phi^T \Phi \alpha}. \quad (6)$$

Then, in order to keep the α values in a reasonable scale and c close to one, we renormalize the α vector at each iteration:

$$\alpha \leftarrow \frac{1}{c}\alpha. \quad (7)$$

3 Multi-Axon Fiber Recovering Metric

One of the problems in the DW-MRI analysis is the definition of a metric that reflects all the aspects of the multi-axon bundle estimation: to estimate the number of fibers, their orientations and their compartment sizes. Several metrics can be used for measure each one of the mentioned aspects, see for instance [14]. However, in the best of our knowledge, there is not a metric that resumes in a scalar value (and thus makes directly comparable different multi-fiber methods) the performance of the fiber estimation methods. Here we propose a new metric for this aim.

First the observed DW-MR signal is an effect of the water diffusion. Such actual diffusion directions are close to some directions represented in the basis functions (see Figure 2). Then, talking about synthetic data, given the observed noisy signal (Panel 2(c)), one wants to recover the fiber information (Panel 2(a)); *i.e.*, we want to minimize the error between the recovered fibers and the actual fibers. Hence, we need to compare the different axon fiber features shown in panels 2(a) and 2(e). Thus, it is necessary to define an error measure that involves those features. However, it is common to use as a quality measure the error between the restored signal, S_{rest} (Panel 2(d)), and the actual noiseless signal, S_{act} (Panel 2(b)):

$$error = \|S_{act} - S_{rest}\|^2. \quad (8)$$

We claim that the correct error metric is the error between the estimated fiber's information in Panel 2(e) and the actual one in Panel 2(a) (one important fiber feature is its orientation). Unfortunately, we have the fiber's information separated. According with [14], we first need to match the fibers which are closer, and then measure three different errors:

- The angular error between the actual and the estimated fiber.
- The magnitude of the difference of the relative diffusion size (size compartment).
- The difference between the number of fibers.

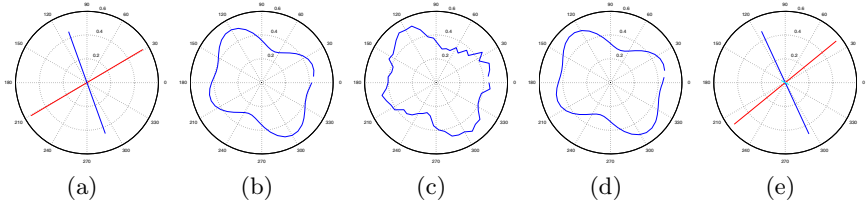


Fig. 2. 2D illustration of the diffusion in a voxel with a crossing axon fiber. (a) Actual fiber orientations, (b) original noiseless signal, (c) noisy observed signal, (d) restored signal and (e) estimated axon fibers.

Nevertheless, it is not clear how to weight and integrate all that information in a single metric.

In the following, we propose a new error measure that considers the fiber’s size compartments as discrete distributions. Our measure takes into account important features of the fibers: the density distribution (shape), the total diffusion and the fiber orientation.

3.1 A Metric for Multi-Fiber Estimations

Here, we propose a metric to measure the error between the recovered multi-fiber solution and the Ground Truth (GT). Before to formally define our metric, it is important to introduce the relevant features of the axon fibers within a voxel.

Let $\beta = \{\beta_1, \beta_2, \dots, \beta_J\}$ and $T = \{T_1, T_2, \dots, T_J\}$ be the actual values (the GT) used for generating synthetic signal by using the equation (1), so that v_i is the largest eigenvector of T_i (the PDD) that is associated with the fiber orientation; then we propose to compute the error between the GT fiber distribution, represented by the vectors β and v , and the estimated one represented by their compartments α and their orientation vectors \tilde{v} with a metric of the form:

$$D_F([\beta, v]; [\alpha, \tilde{v}]) = D_o([\beta, v]; [\alpha, \tilde{v}]) + D_v(\beta; \alpha) + D_s(\beta; \alpha). \quad (9)$$

Each term in (9) is explained in the following. Note that the number of elements in β and α is different. The same case occurs with v and \tilde{v} .

Orientation Term: D_o . Given two fibers, with PDDs v_i and \tilde{v}_j , we define the angle $[0, \pi]$ between the PDDs as orientation distance:

$$d(v_i, \tilde{v}_j) = \arccos(v_i \cdot \tilde{v}_j). \quad (10)$$

From now on, we denote $d_{ij} = d(v_i, \tilde{v}_j)$. Then, we can see d_{ij} as the cost (measure of error) of saying that \tilde{v}_j is the orientation of a single fiber when the real orientation is v_i . In our problem we can have several fibers, so we need to match the corresponding compartment in order to find the cost. Such a match must be the one that computes the minimum cost; *i.e.* we need to match the closest

fibers. Hence, the computation of this metric is formulated as the solution of the transportation problem:

$$D_o([\beta, v]; [\alpha, \tilde{v}]) = \min_x \frac{\sum_{i,j} d_{ij} x_{ij}}{\sum_{i,j} x_{ij}}$$

subject to

$$\sum_j x_{ij} \leq \beta_i,$$

$$\sum_i x_{ij} \leq \alpha_j \quad (11)$$

$$x_{ij} \geq 0$$

$$\sum_{i,j} x_{ij} = \min \left\{ \sum_i \beta_i, \sum_j \alpha_j \right\}.$$

In this manner, x_{ij} denotes the transportation flows and represents the amount transported from the i -th supply to the j -th demand.

The system of equations (11) is known as the Earth Mover's Distance (EMD). The EMD is a measure which evaluates the dissimilarity between two multi-dimensional distributions in some feature space by using a distance measure between single features. This distance is defined as the minimal cost that must be paid to transform one distributions into the other [16,17].

Total Diffusion Term: D_v . The term D_v measures the distance between the total diffusion of the estimated fiber and the GT, we use:

$$D_v(\beta; \alpha) = c_v \left| \sum_i \beta_i - \sum_j \alpha_j \right|. \quad (12)$$

where c_v is a positive scalar that weights the contribution of the term. In order to make comparable the contribution of this term w.r.t the previous term, D_o , a choice for c_v is:

$$c_v = \max_{i,j} \{d_{ij}\}. \quad (13)$$

where d is defined in (10). The combination of D_o and D_v using this particular selection for c_v was previously reported by Pele and Werman for comparing unnormalized image color histograms [18].

Sparsity Term: D_s . The last term D_s measures the difference between the sparsity of the the GT and the estimation. Let

$$N(\beta) = \frac{\beta}{\sum_k \beta_k} \quad (14)$$

be the normalization operator. Thus, we see $N(\beta)$ and $N(\alpha)$ as discrete density probability functions (for an orientation distribution) and compare their sparsity by comparing their Shannon's entropy:

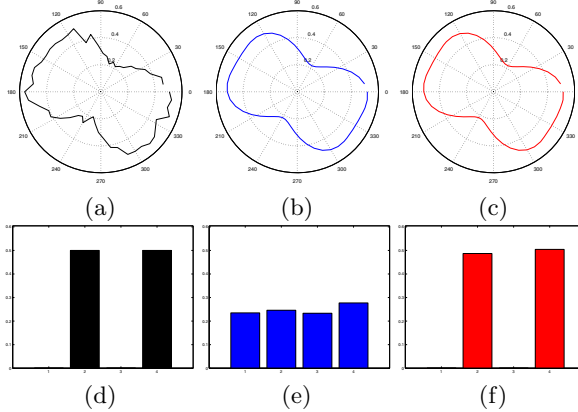


Fig. 3. 2D example of the estimation of β coefficients by using the equations (4) and (5). The noisy signal (a) was built with the β coefficients showed in (d), noise with Rician distribution was added. (b) and (e) show the restored signal and the computed basis coefficients with (4). (c) and (f) show the restored signal and the computed basis coefficients by our proposal in (5). Note that, even the recovered signals are similar, the new methodology does not overestimate the number of axon bundles.

$$D_s(\beta; \alpha) = c_s \left| \sum_i N(\beta)_i \log N(\beta)_i - \sum_j N(\alpha)_j \log N(\alpha)_j \right| \quad (15)$$

where c_s is a parameter that keeps the value of D_s comparable with D_o and D_v ; we use:

$$c_s = \max \left\{ \sum_i \beta_i, \sum_j \alpha_j \right\}. \quad (16)$$

4 Experiments and Results

In this section we present the results obtained on synthetic DW-MRI and real DW-MRI data.

4.1 Synthetic 2D Example

First, we show a 2D example to illustrate the methodology. Figure 3 shows the results obtained by solving the equations (4) and our proposal in (5). Note that in this case the observed signal was built as a linear combination of the signals basis functions (this is not always the case, but allows to synthetically illustrate our proposal) and noise with Rician distribution was added. As mentioned before, the values of the size compartment are visualized as a discrete distribution. One can see that solution by using (4) gives more signal basis than the solution by solving (5) and both restored signals are similar. The overestimation problem on formulation (4) is because the signal fitting does not penalize the number of

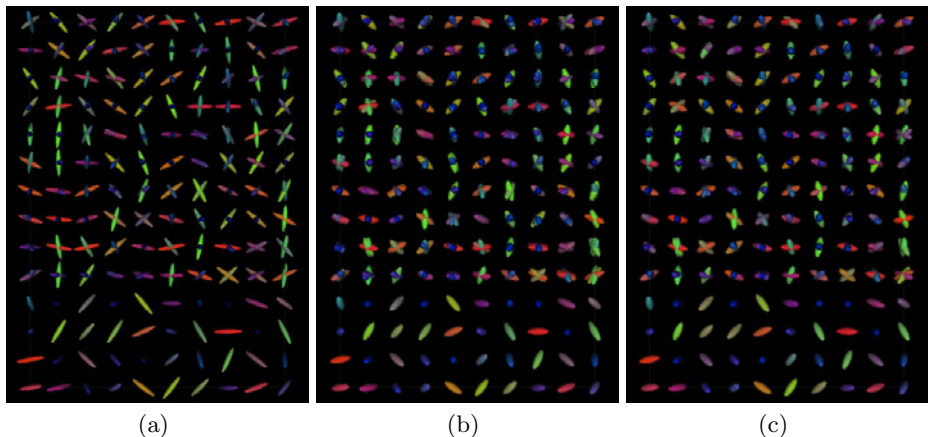


Fig. 4. Results on synthetic data with $SNR = 10$. (a) actual tensors, (b) estimation using the equation (4) and (c) estimation using our proposal in equation (5).

coefficients and only minimizes the difference between signals, as opposite that in our formulation.

Moreover, the proposed error measure D_F detects the difference between the two solutions: the error with respect to the GT for the solution in Panel 3(e) and 3(f) are 0.5592 and 0.0137, respectively. Thus, our metric indicates that the proposed sparse solution is closer to the GT. Note how our metric allows to detect differences and similarities between all the features of axon multi-fiber solutions by the evaluation of a single metric.

4.2 Synthetic 3D Example

In this experiment, we used 3D synthetic data with the following description:

- **Synthetic data.** The DW-MRI signal was synthesized from the GMM (1). The DT principal eigenvalue was set to $1 \times 10^{-3} \text{ mm}^2 / \text{s}$ and the second and third tensor eigenvalues were $2.22 \times 10^{-4} \text{ mm}^2 / \text{s}$, $FA = 0.74$. The above values were taken from a sample of tensors observed in the brain data from a healthy volunteer. The tensors were randomly rotated to generate a composed field of 140 voxels. Rician noise was added to each measurement to produce a low Signal to Noise Ratio (SNR) equal to 10, which is typically found on clinical settings.

Now, in Figure 4 we show the results obtained by using the synthetic data. The Panel 4(a) shows the GT. Note that the synthetic data represent two cases: voxels with crossing fibers and voxels with single fiber. We can see that formulation (5) gives a closer solutions to the GT w.r.t. the formulation in (4). Moreover, Figure 5 shows the number of tensors in each voxel for the GT and the two formulations. One can see that the number of tensors computed by formulation in (5) is closer to the GT, i.e. it is more sparse and eliminate the over estimation

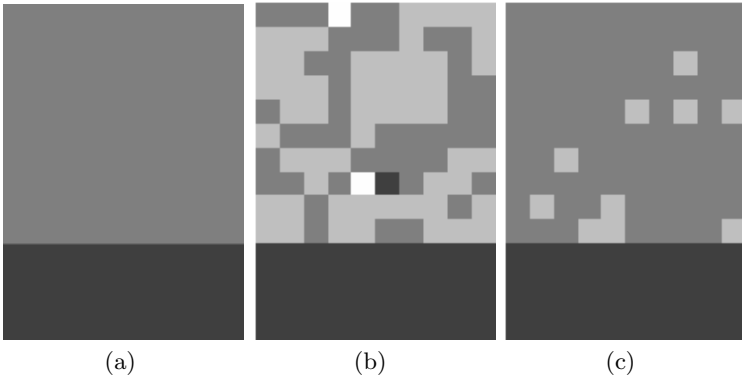


Fig. 5. Number of tensor in each voxel for experiment in Figure 4. We show the number of tensors coded as grey scale (dark = 1, dark grey = 2, light grey = 3 and white = 4). (a) Ground truth of synthetic data, (b) estimation by equation (4) and (c) estimation using our proposal in equation (5). The overestimation problem is improved in our formulation.

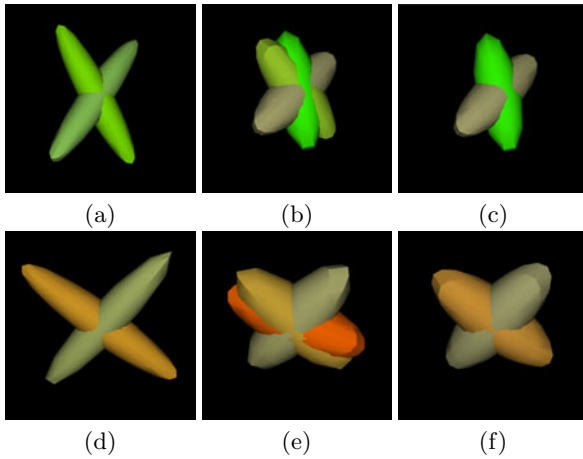


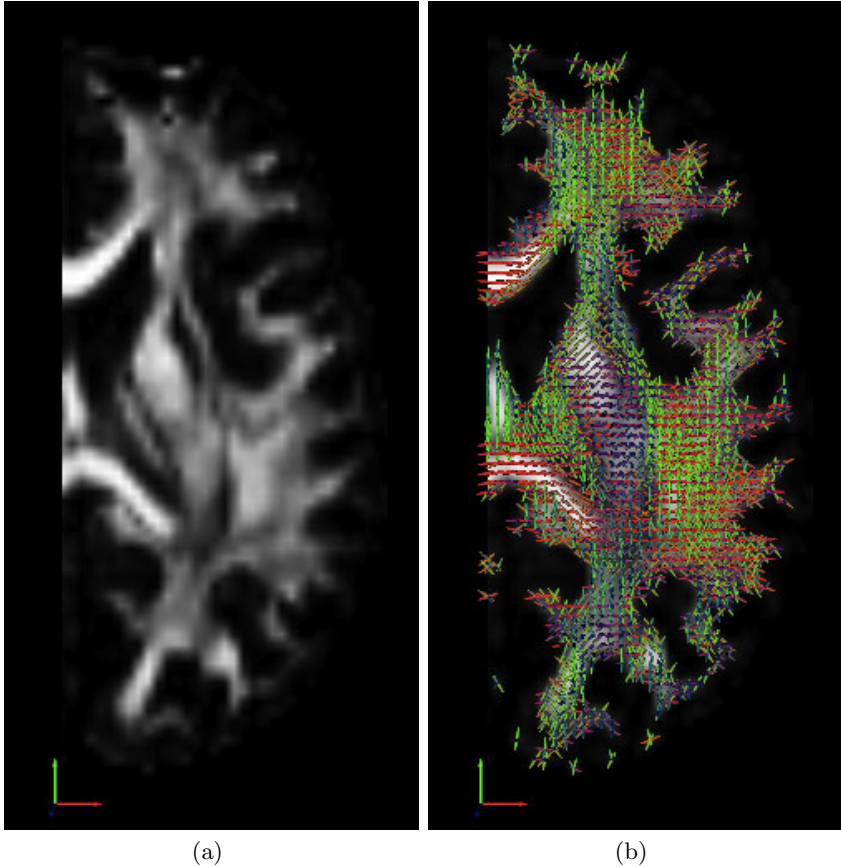
Fig. 6. Zoom of some voxels of the synthetic data of the Figure 4. (a) and (d) are the GT. (b) and (e) are the results by solving (4). (c) and (f) are the result by using our proposal in (5).

problem of formulation in (4). We select two representative voxels in order to show details of the computed solutions. Figure 6 shows the solutions associated to the two tested formulations.

The Table 1 shows the D_F mean error values, the D_F error variance and the percentage of voxels that had less error w.r.t the other method for formulations in (4) and (5). We can see that the mean error and the error variance are lower for our proposal. Also, in a 91.43% of the cases, the new formulation has a lower error.

Table 1. D_F mean and D_F variance of the 3D synthetic data by using (4) and (5)

Method	mean error	variance error	lowest error (percentage)
Original formulation	0.7624	2.0e-03	8.57%
New formulation	0.2995	8.5e-04	91.43%

**Fig. 7.** Results *In vivo* Brain Human Data. (a) White matter struct by showing the Fractional Anisotropy map, (b) fiber estimations using the equation 5.

4.3 Real Data

The last experiment was performed on *in vivo* Brain Human data, these data have the following features:

- ***In vivo* Brain Human Data:** A single healthy volunteer was scanned on a Siemens Trio 3T scanner with 12 channel coil. Acquisition parameters: single-shot echo-planar imaging, five images for $b=0$ s/mm, 64 DW images with unique, isotropically distributed orientations ($b=1000$ s/mm²),

TR=6700 ms, TE=85 ms, 90° flip angle, voxel dimensions equal to $2 \times 2 \times 2$ mm³. The approximated SNR = 26.

The Figure 7 shows the results obtained by using the proposal in brain data. We select a region of interest, such that the center of the image shows crossing axon fibers within the right *superior longitudinal fasciculus*. The Panel 7(a) shows the white matter struct by showing the Fractional Anisotropy (FA) map [19]. The Panel 7(b) shows the estimated multi-tensors. One can see that our sparse solution still captures the complex multi-fiber brain showing several well-known crossing fibers and bifurcations.

5 Conclusions

In this article, we presented a new formulation for Diffusion Basis Functions fitting which improves the original formulation by reducing the overestimation effect on number of estimated fibers due to signal noise. The new formulation adds the Tikhonov's regularization term. The improvements of the proposed method was showed on synthetic and real human brain images.

In addition, we presented a new metric to measure the error between two multi-fiber solutions, we note that it differentiates them properly. This metric models the size compartments of the fibers like piles of earth. Also, the proposed measure takes into account the angular error, the number of the fibers and its distribution in one only formulation based on the Earth Mover's Distance.

References

1. Buxton, R.B.: Introduction to Functional Magnetic Resonance Imaging: Principles and Techniques, 1st edn. Cambridge University Press (2002)
2. Poldrack, R.A.: A structural basis for developmental dyslexia: Evidence from diffusion tensor imaging. In: Wolf, M. (ed.) Dyslexia, Fluency, and the Brain, pp. 213–233. York Press (2001)
3. Basser, P.J., Mattiello, J., Lebihan, D.: MR Diffusion Tensor Spectroscopy and Imaging. *Biophysical Journal* 66, 259–267 (1994)
4. Basser, P.J., Pierpaoli, C.: Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI. *J. Magn. Reson. B* 111 (1996)
5. Aranda, R., Rivera, M., Ramírez-Manzanares, A., Ashtari, M., Gee, J.C.: Massive Particles for Brain Tractography. In: Sidorov, G., Hernández Aguirre, A., Reyes García, C.A. (eds.) MICAI 2010, Part I. LNCS, vol. 6437, pp. 446–457. Springer, Heidelberg (2010)
6. Stejskal, E.O.: Use of Spin Echoes in a Pulsed Magnetic-Field Gradient to Study Anisotropic, Restricted Diffusion and Flow. *The Journal of Chemical Physics* 43, 3597–3603 (1965)
7. Ramírez-Manzanares, A., Rivera, M.: Basis tensor decomposition for restoring intra-voxel structure and stochastic walks for inferring brain connectivity in DT-MRI. *Int. Journ. of Comp. Vis.* 69, 77–92 (2006)
8. Bergmann, O., Kindlmann, G., Peled, S., Westin, C.F.: Two-tensor fiber tractography. In: IEEE 2007 International Symposium on Biomedical Imaging (ISBI), Washington D.C. (2007)

9. Malcolm, J.G., Michailovich, O., Bouix, S., Westin, C.F., Shenton, M.E., Rathi, Y.: A filtered approach to neural tractography using the watson directional function. *Medical Image Analysis* 14, 58–69 (2010)
10. Tuch, D.S., Reese, T.G., Wiegell, M.R., Makris, N., Belliveau, J.W., Wedeen, V.J.: High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity. *Magn. Reson. Med.* 48, 577–582 (2002)
11. Ramírez-Manzanares, A., Rivera, M., Vemuri, B.C., Carney, P., Mareci, T.: Diffusion basis functions decomposition for estimating white matter intravoxel fiber geometry. *IEEE Trans. Med. Imag.* 26, 1091–1102 (2007)
12. Ramírez-Manzanares, A., Rivera, M.: Basis pursuit based algorithm for intra-voxel recovering information in DW-MRI. In: *Proc. IEEE Sixth Mexican International Conference on Computer Science (ENC 2005)*, Puebla, México, pp. 152–157 (2005)
13. Jian, B., Vemuri, B.: A unified computational framework for deconvolution to reconstruct multiple fibers from diffusion weighted MRI. *IEEE Trans. Med. Imaging* (2007)
14. Ramírez-Manzanares, A., Cook, P.A., Gee, J.C.: A comparison of methods for recovering intra-voxel white matter fiber architecture from clinical diffusion imaging scans. *Med. Image Comput. Comput. Assist. Interv.*, 305–312 (2008)
15. Nocedal, J., Wright, S.J.: *Numerical optimization*. Springer, Heidelberg (1999)
16. Rubner, Y., Tomasi, C., Guibas, L.J.: The Earth Mover’s Distance as a Metric for Image Retrieval. *International Journal of Computer Vision* 40, 99–121 (2000)
17. Pele, O., Werman, M.: Fast and Robust Earth Mover’s Distances. In: *2009 IEEE 12th International Conference on Computer Vision*, pp. 460–467. IEEE (2009)
18. Pele, O., Werman, M.: A linear Time Histogram Metric for Improved Sift Matching. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part III*. LNCS, vol. 5304, pp. 495–508. Springer, Heidelberg (2008)
19. Le Bihan, D., Mangin, J.F., Poupon, C., Clark, C.A., Pappata, S., Molko, N., Chabriat, H.: Diffusion tensor imaging: Concepts and applications. *J. Magn. Reson. Imaging* 13, 534–546 (2001)

An Adaptive Motion Data Storage Reduction Method for Temporal Predictor

Ruobing Zou, Oscar C. Au, Lin Sun, Sijin Li, and Wei Dai

Department of Electronic and Computer Engineering,
Hong Kong University of Science and Technology,
Clear Water Bay, Kowloon, Hong Kong
{zouruobing, eeau, lsunece, sliae, weidai}@ust.hk

Abstract. In the state-of-art video coding standard HEVC, temporal motion vector (MV) predictor is adopted in order to improve coding efficiency. However, motion vector information in reference frames, which is used by temporal MV predictor, takes significant amount of bits in memory storage. Therefore motion data needs to be compressed before storing into buffer. In this paper we propose an adaptive motion data storage reduction method. First, it divides the current 16x16 block in the reference frame into four partitions. One MV is sampled from each partition and all sampled MVs form a MV candidate set. Then it judges if one or two MVs should be stored into the MV buffer by checking the maximum distance between any two of the MVs in the candidate set. If the maximum distance is greater than a certain threshold, the motion data of the two MVs that have maximum distance are put into memory; otherwise the motion data of the upper left block is stored. The basic goal of the proposed method is to improve the accuracy of temporal MV predictor at the same time reducing motion data memory size. Simulation results show that compared to the original HEVC MV memory compression method in the 4th JCT-VC meeting, the proposed scheme achieves a coding gain of 0.5%~0.6%; and the memory size is reduced by more than 87.5% comparing to without using motion data compression.

Keywords: HEVC, MV data storage, MV compression, temporal predictor.

1 Introduction

The Motion Pictures Experts Group (MPEG) and the International Telecommunications Union's Telecommunication Standardization Sector (ITU-T) recently have formed a joint collaborative team on video coding (JCT-VC). The JCT-VC is aiming at designing a next generation video coding standard. This standard, currently named as High Efficiency Video Coding (HEVC), targets substantial coding efficiency improvement compared to state-of-the-art video coding standards such as MPEG-4 and H.264/AVC [1]. In order to evaluate different coding techniques, a software platform called HEVC Test Model (HM) is developed, which contains selected coding tools. In HEVC, the MVs of reference frames are used in the prediction process of a current

frame as temporal MV candidates, such as in Advanced Motion Vector Prediction (AMVP) and in PU-level Merge mode [2]. Therefore, temporal information needs to be stored in a buffer, which occupies a portion of memory size. In order to save the memory for storing MV data of the reference frames, a MV memory compression method was adopted by the 4th JCT-VC meeting and written into the HEVC working draft [3]. However, as reported in [2], this MV memory compression method results in a coding loss of 0.6%~0.8% in terms of BD Bit Rate (BD-rate) [4], compared with turning off MV compression tool. JCT-VC recently formed a new Core Experiment to study the problem of motion data storage. Under this background, we propose a novel motion data storage reduction scheme which can reduce the memory size while improving coding efficiency.

The remainder of the paper is organized as follows. In Section 2, we generally describe inter coding tools in HEVC, to which temporal predictor is applied, and then we explain how MV data is compressed and stored. In Section 3, the proposed method is introduced. Results on HEVC Test Model HM2.0 software are presented in section 4. Conclusions are drawn in section 5.

2 Motion Data Storage Reduction

2.1 Prediction Unit (PU) Merge Mode and Advanced Motion Vector Prediction (AMVP)

In terms of HEVC, two different motion information prediction schemes exist in the inter frame coding. One of the motion information prediction schemes is merge mode. Merge mode is performed on PU level, where PU is the basic unit used for carrying the information that is related to the prediction processes. The merge mode infers that the inter prediction direction and the reference frame index of the current PU are the same as one of the neighbouring blocks and uses their MVs as predictors. Merge candidates are four spatial predictors and one temporal predictor, as shown in Fig. 1(a). Temporal predictor is the MV of the co-located PU partition, labeled “Col” in Fig. 1 (a) and (b). Derivation of the temporal predictor will be explained in detail later.

The other prediction scheme is AMVP which uses MVs from the neighbouring PUs as MV predictors [5]. There are totally three possible MV predictors: two spatial predictors “Left” and “Above”, as well as a temporal predictor. AMVP “Left” predictor is the first 4x4 neighbour found along the left edge of the current PU, while the “Above” predictor is the first available 4x4 neighbour found along the top edge, as illustrated by Fig. 1(b). We will further describe the temporal MV predictor in below.

2.2 Temporal Motion Vector Predictor

In both the merge mode and AMVP, the MVs of reference frames are used as the temporal predictors to improve coding efficiency. As determined in the 4th HEVC

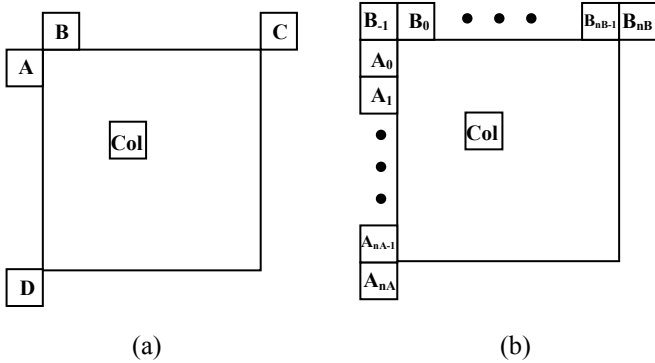


Fig. 1. Predictors for Merge mode and AMVP. (a) Neighboring blocks checked by merge mode, including the temporal predictor (Col). (b) Search range for left predictor (A_0 to A_{nA}), above predictor (B_{-1} to B_{nB}) and co-located temporal predictor (Col).

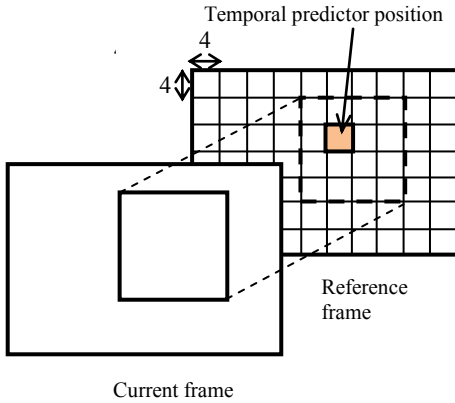


Fig. 2. Position of the center temporal predictor

meeting, the temporal predictor is found by mapping the upper left position of the center of the current partition to a co-located block in reference frame, and the MV of the corresponding 4x4 block in the co-located block is used as a temporal MV predictor [6]. By using the temporal MV predictor, it shows that the average 2.2% of BD-rate saving is achieved for random access and low delay test conditions [7].

2.3 Reduced Resolution Storage of Motion Vector Data

Despite of the BD-rate saving brought by the temporal MV predictor, the usage of the temporal MV leads to the need of additional memory requirement in inter prediction because the MV data information of the reference frames needs to be stored. Without

compression, each 4x4 block has its own MV, which needs significant buffer size, considering the granularity of motion representation and considering that there are two vectors per block for B slice. To give an example, it is currently estimated the buffer size to be approximately 26Mbits/frame for 4kx2x application, though this size will ultimately depend on the precision and maximum MVs supported [8]. Large amount of MVs to be stored causes the internal memory size to increase, which may result in increasing hardware cost and power consumption [9].

In order to reduce the motion data storage, a memory compression method in [8] was adopted at 4th JCT-VC meeting, in which a lower resolution MV buffer is used. It means that within a predefined 16x16 block, only one MV in the top left 4x4 block is saved into the buffer, instead of sixteen different MVs, as illustrated by Fig. 3 (a) and (b). Indeed not only MV fields but also reference frame indices and modes of the reference frame (inter/intra) are used in temporal MV prediction and thus they are compressed in the same method as well. The saved MV is used as a representative MV of the current 16x16 block, and when a reference frame is referred as a co-located block for the temporal predictor, the representative MV is read from the buffer and then assigned to all 4x4 blocks within the current 16x16 block (Fig. 3(c)). Therefore, both the required memory size and the memory access bandwidth for the temporal MV data are reduced to 1/16.

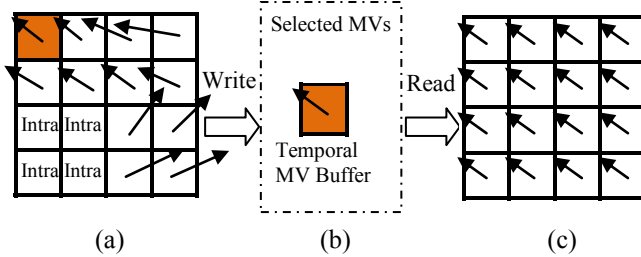


Fig. 3. Illustration of the MV compression in HEVC. (a) Motion vectors before compressed. (b) After MV compression. (c) In referring as co-located block for temporal predictor. [10]

3 Proposed Scheme

In this section, we first analyze the HEVC MV compression scheme introduced before and show its disadvantage in some cases. Then we propose an adaptive MV compression scheme.

3.1 Analysis of the Limitations of Storing Upper Left Block MV Information

To illustrate the limitation of the original memory compression method in HEVC, we examine the sixteen MVs in a current 16x16 block before compression. Suppose that

the current block lies inside an object which has translational motion. Under this situation nearly all 4x4 blocks have similar MV, therefore, MV of the upper left 4x4 block (denoted by mv_0) works well as a representative of the current block. However, the current scheme ignores that there could be two or more objects located inside a 16x16 block. If these two objects have different movement speed and directions, then simply assigning mv_0 to all blocks may result in a large deviation from a block's real MV, consequently causing reliability degradation of the temporal predictor and introducing coding loss.

3.2 Proposed Motion Data Storage Reduction Scheme

In order to reduce the coding loss, it is natural to consider storing more MVs into memory. However, increasing data in the storage inevitably leads to lower compression rate. Therefore it is crucial to achieve a good trade-off between these two factors. Based on this idea, we propose a novel method which reduces the MV buffer size by adaptively choosing to store either one or two MVs from the sixteen MVs of the current block. The details of the proposed algorithm can be depicted as follows.

The proposed scheme first divides the current 16x16 block into four partitions. As depicted in Fig. 4, partition 1 to 4 is shown with horizontal shading, down diagonal shading, up diagonal shading and vertical shading separately. Each partition contains four 4x4 blocks and thus it has four MVs. We denote the MV of block n as mv_n , $\forall n \in \{0, \dots, 16\}$. The MV set of the partition k is denoted as N_k , $\forall k \in \{1, \dots, 4\}$. If a block's MV is not available, when the block is in intra mode or out of the slice boundary, then its MV is set to be (0,0).

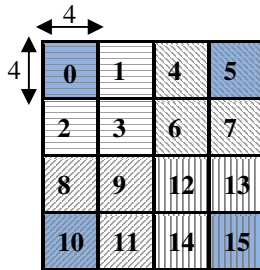


Fig. 4. Partitions and sampled MVs in proposed scheme. Partition 1 contains block 0, 1, 2 and 3; partition 2 contains block 4, 5, 6 and 7; partition 3 contains block 8, 9, 10 and 11; partition 4 contains block 12, 13, 14 and 15.

Following the partitioning, a sampling is conducted on every MV set N_k . One MV is sampled out of four and becomes a member of set P , which is a set of MV candidates that might be put into the buffer. All of the sampled MVs are on the corner of the current 16x16 block, and thus the sampled MV of partition k is $mv_{5(k-1)}$, $k \in$

$\{1, \dots, 4\}$, as shown in dark colour shading in Fig. 4. In this way, the MV candidate set $P = \{mv_0, mv_5, mv_{10}, mv_{15}\}$ is obtained. The reason why blocks on the corner are chosen is because these blocks are far from each other thus they are relatively with low correlation. This is particularly true when two or more objects exist, whose MVs pointing to different directions. It ensures that different movement conditions are taken into consideration simultaneously.

With the candidate set P , Euclidean distance among any two MVs in P is calculated as follows:

$$d(i, j) = \|mv_i - mv_j\|, \quad (1)$$

Where $mv_i, mv_j \in P$, $i \neq j$ and $i, j \in \{0, 5, 10, 15\}$. We calculate the Euclidean distance because x-component and y-component of vectors are considered together. The maximum distance d_{max} is used in checking whether to store one or two set(s) of MV data as follows. Suppose $mv_a, mv_b \in P$ satisfy:

$$d(a, b) = \|mv_a - mv_b\| = d_{max},$$

Where mv_a is the sampled MV of partition A, mv_b is the sampled MV of partition B, $a \neq b$ and $A, B \in \{1, \dots, 4\}$. The basic rule can be described as:

$$\text{MVs stored} = \begin{cases} mv_0, & \text{if } d_{max} \leq T, \\ mv_a \text{ and } mv_b, & \text{if } d_{max} > T, \end{cases} \quad (2)$$

If d_{max} is smaller than a certain threshold T , MVs in P are considered to be similar. In this case, using only one MV can well represent all MVs in the entire 16x16 block. Hence, MV information of the upper left block, namely mv_0 , frame indices and modes, is stored into the MV buffer, which is basically the same as the memory compression method in HEVC. Besides, a one-bit MV_num_flag is set to 1 as a mark of putting only one MV into buffer.

On the other hand, if d_{max} is greater than T , it means that at least one partition has MV deviating from others, so that the algorithm adaptively stores both mv_a and mv_b into the MV buffer, along with their frame indices and modes. Similarly, the MV_num_flag is set to 0 indicating that two MVs are selected. In addition, we store a one-bit flag for every partition so as to indicate that it uses either mv_a or mv_b as its representative. Specifically, we set the flag of partition A (flag_A) to be 1, indicating that it uses mv_a as representative; and flag of partition B (flag_B) to be 0. With regard to the other two partitions, each of them also needs to select a representative MV from the stored MVs in the following method. Suppose these two partitions are partition C and D, with sampled MV mv_c and mv_d , where $mv_c, mv_d \in P$, $a \neq b \neq c \neq d$ and $C, D \in \{1 \dots 4\}$. For each of mv_c and mv_d , we check its distance to mv_a and mv_b . Take partition C as an example, the representative MV (rMV) of partition C equals to:

$$rMV = \begin{cases} mv_a, & \text{if } d(c,a) < d(c,b), \\ mv_b, & \text{if } d(c,a) \geq d(c,b), \end{cases} \quad (3)$$

Meanwhile, a flag is set to indicate which stored MV is used as representative of partition C. To be specific, if mv_a is selected as the representative, the flag of partition C ($flag_C$) is set to be 1; otherwise the flag equals to 0. The same process is performed on mv_d according to equation (3). A flowchart of the whole process of motion data compression is shown in Fig. 6.

So far, we have introduced how motion data is adaptively selected and compressed into the MV buffer. After that, when the 16x16 block is referred as co-located block in temporal prediction, the MV(s) stored will be read from buffer. If two MVs are stored, they are assigned to the partitions as representative according to the flags we previously stored; otherwise the single stored MV is assigned to all partitions. Fig. 5 (a) ~ (e) shows an example of how to store MV data adaptively based on different situation and how to assign MV(s) that has/have been read from buffer.

As a summary, the memory size of the proposed method is bigger than HEVC's method, whose memory size is reduced to 6.25% and smaller than constantly storing two MVs, whose memory size is 12.5%. So the total reduction of the memory size is more than 87.5% comparing to not using MV compression.

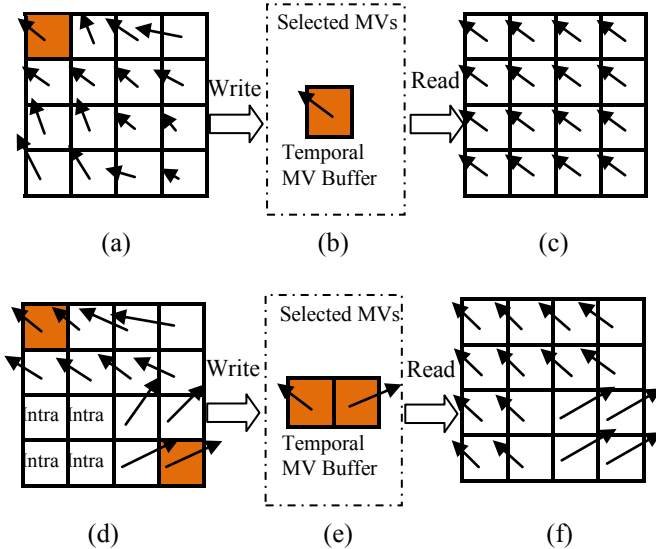


Fig. 5. Illustration of proposed motion data compression scheme. (a) ~ (c) Illustrates the case of storing one MV, (d) ~ (f) illustrates the case of storing two MVs. (a) (d) Motion vectors before compressed. (b) (e) After motion data compression. (c) (f) In referring as co-located block for temporal predictor.

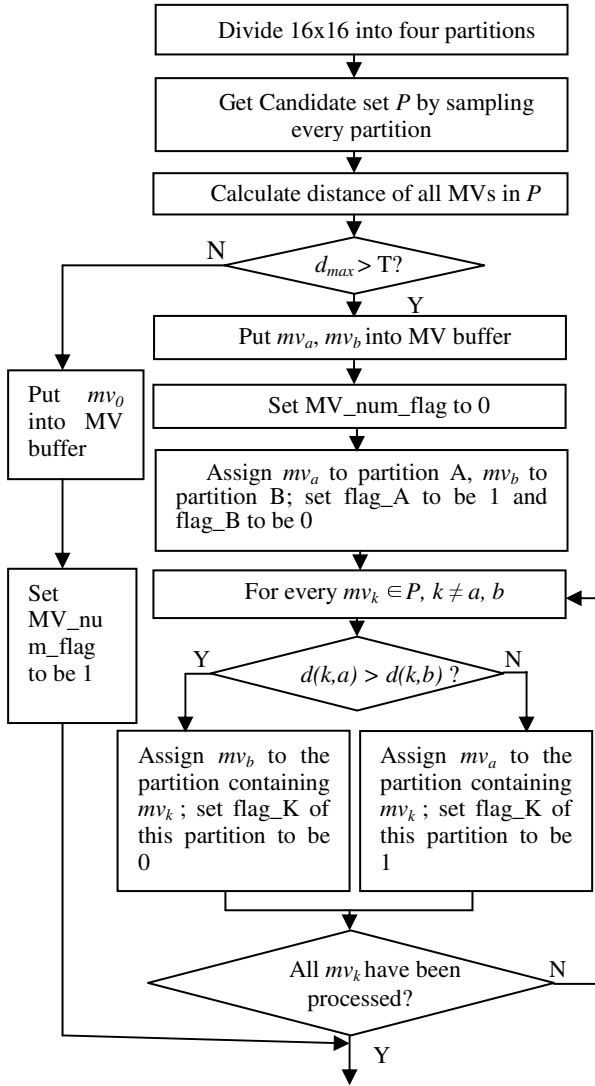


Fig. 6. Detailed process of motion data compression

4 Experimental Results

To measure the performance of the proposed motion data compression technique, we compare the performance of proposed scheme with HEVC MV compression scheme described in Section 2.

4.1 Test Conditions

We execute experiments on HM software version 2.0, which is the latest version of HEVC Test Model created following the decisions taken at the 4th meeting of JCT-VC in Daegu in January 2011. The platform of experiment is Intel i7 CPU 860 @2.80GHz with 16GB RAM. Video sequences to be tested are provided by JCT-VC, including five classes of sequence: Class A, Class B, Class C, Class D and Class E, with resolution 2560x1600, 1920x1080, 832x480, 416x240 and 1280x720 pixels. For each video sequence four quantization parameter values are used: 22, 27, 32 and 37.

Our experiment is executed with common test conditions in HEVC to make sure of conducting experiments in a well-defined environment and to ease the comparison of the outcome of experiments [11]. Four common test conditions are used in inter prediction, reflecting a combination of high efficiency (HE) and low complexity (LC), and of random-access (RA), and low-delay (LD) settings:

- Random access, high efficiency (RA-HE)
- Random access, low complexity (RA-LC)
- Low delay, high efficiency (LD-HE)
- Low delay, low complexity (LD-LC)

Noted that when testing MV compression tools, low-delay configurations should be skipped for class A, and random-access configurations should be skipped for class E. Table 1 gives the specifications of sequences used for random-access, and low-delay conditions.

Table 1. Basic information of the test sequences

Class	Sequence name	Frame count	Frame rate	RA	LD
A (4k)	Traffic	150	30fps	Y	N/A
	PeopleOnStreet	150	30fps	Y	N/A
B (1080p)	ParkScene	240	24fps	Y	Y
	Cactus	500	50fps	Y	Y
	Kimono	240	24fps	Y	Y
	BasketballDrive	500	50fps	Y	Y
	BQTerrace	600	60fps	Y	Y
C (WVGA)	RaceHorses	300	30fps	Y	Y
	BasketballDrill	500	50fps	Y	Y
C (WVGA)	BQMall	600	60fps	Y	Y
	PartyScene	500	50fps	Y	Y
D (WQVGA)	RaceHorses	300	30fps	Y	Y
	BasketballPass	500	50fps	Y	Y
	BQSquare	600	60fps	Y	Y
	BlowingBubbles	500	50fps	Y	Y
E (WQVGA)	Vidyo1	600	60fps	N/A	Y
	Vidyo3	600	60fps	N/A	Y
	Vidyo4	600	60fps	N/A	Y

Tests are executed by separately turning on a tool of HEVC MV compression, and turning on a tool of the proposed method. When implementing the latter experiment, we use the square of MV distance instead of MV distance because calculating square root increases complexity. As for choosing the threshold T, when the distance between two vectors is more than the maximum distance in a 4x4 block, we take them as deviating from each other and store both of them. So the square of T is set to be 32, which is the largest distance within a 4x4 block. Also, noted that the HEVC working draft stated a bug in HM 2.0 [12], however, the current software does not exactly implement this. We fixed this bug in our experiment, i.e. reference index and mode is also decimated by taking the values that apply to the top left pixel position of each 16x16 block, as motion vector field does.

Table 2. Coding performance of the proposed scheme

Class	RA-HE			RA-LC		
	BD-rate			BD-rate		
	Y	U	V	Y	U	V
A	-0.7	-0.6	-0.5	-0.6	-0.6	-0.6
B	-0.2	-0.2	-0.2	-0.2	-0.1	-0.1
C	-0.4	-0.4	-0.4	-0.4	-0.3	-0.4
D	-0.6	-0.6	-0.6	-0.8	-0.6	-0.6
E						
All	-0.5	-0.5	-0.4	-0.5	-0.3	-0.4
Enc Time[%]	102			101		
Dec Time[%]	101			100		

Class	LD-HE			LD-LC		
	BD-rate			BD-rate		
	Y	U	V	Y	U	V
A						
B	-0.3	-0.2	-0.3	-0.4	-0.4	-0.3
C	-0.6	-0.4	-0.6	-0.7	-0.7	-0.7
D	-0.6	-0.3	-0.8	-0.6	-0.6	-0.5
E	-0.7	-0.6	-0.2	-0.5	-0.5	-0.9
All	-0.6	-0.4	-0.5	-0.6	-0.4	-0.7
Enc Time[%]	102			101		
Dec Time[%]	101			100		

4.2 Results and Analysis

After the performance of MV compression method in HEVC and that of the proposed method are both tested, their test results are compared in terms of BD-rate [4], which is computed by Excel sheet provided by JCT-VC. The negative value of BD-rate

indicates a coding gain of the proposed method, comparing with the HEVC MV compression method in HM 2.0.

Experimental results are shown in Table 2; it is observed that the proposed method shows 0.5%, 0.5%, 0.6% and 0.6% BD-rate gain, which indicates consistently improvement in all test conditions. In random access cases, the proposed method shows 0.5% of BD-rate saving than the HEVC method for high efficiency, and 0.5% of BD-rate saving than HEVC method for low complexity on average. While in low delay case, the proposed method saves 0.6% of BD-rate saving for high efficiency and 0.6% for low complexity on average. Meanwhile, the difference in encoding and decoding time is relatively small.

In order to show the gap between with and without MV compression, we also cite the experimental results in [2], which tested on the HM2.0 software in two conditions: turning on and turning off with HEVC MV compression tool. The results show that without MV compression, it achieves 0.6%, 0.7%, 0.8% and 0.8% BD-rate gain for RA-HE, RA-LC, LD-HE and LD-LC configuration, which means that turning on the tool brings a coding loss of 0.6%, 0.7%, 0.8% and 0.8% BD-rate. It can be seen from all above results that the performance loss taken by the MV compression in HEVC can be mostly compensated by using the proposed MV data storage scheme. The proposed motion data storage reduction scheme keeps the loss to 0.1%~0.2%.

The advantage of the proposed method over the original HEVC MV compression method is that it adds only one MV to temporal MV buffer adaptively in certain cases, but it in turn brings relatively high coding gain than the original method. The gain comes mainly from the fact that our method effectively reduced the inaccuracy of the temporal predictor due to MV compression.

5 Conclusion

In this paper, a novel motion data storage reduction method is proposed. In the proposed method, for every 16x16 block in a reference frame, it judges if an additional memory area should be used for motion data compression by checking the maximum MV distance within the current block, and then it adaptively selects one or two MVs to store into the MV buffer. By doing so, the proposed method avoids a waste of memory caused by always storing two MVs; at the same time, it also reduced the inaccuracy of temporal MV predictor caused by constantly storing one MV. Compared with the memory compression method in HEVC, coding gain can be observed in terms of BD-rate saving, which is 0.5%, 0.5%, 0.6% and 0.6% for RA-HE, RA-LC, LD-HE and LD-LC configuration. In other words, the proposed scheme improves coding efficiency for various video sequences than the MV compression method in HEVC, while reducing motion data memory size by more than 87.5% comparing to without motion data compression. In the recently released software HM 3.0, the position of temporal MV predictor was changed to the right corner of the current PU. Our further research will include finding an effective MV data storage method in the new condition.

References

1. Segall, A., Zhao, J., Yamamoto, T.: Parallel Intra Prediction for Video Coding. In: Picture Coding Symposium, Nagoya, Japan, pp. 310–313 (2010)
2. Guo, X., Lin, J., Huang, Y.W., Lei, S.: Motion Vector Decimation for Temporal Prediction. In: JCT-VC 5th Meeting, JCTVC-E092, Geneva (2011)
3. WD2: Working Draft 2 of High-Efficiency Video Coding. In: JCT-VC 4th Meeting, CTVC-D503, Daegu, Korea (2011)
4. Bjontegaard, G.: Calculation of Average PSNR Differences Between RD Curves. In: ITU-T SC16/Q6, VCEG-M33, Austin, USA (2004)
5. Yeo, C., Tan, Y.H., Li, Z.: Simplified AMVP Candidate Derivation For Inter and Merge Modes. In: JCT-VC 5th Meeting, JCTVC-E101, Geneva (2011)
6. Park, S., Park, J., Jeon, B.: Modifications of Temporal MV Compression and Temporal MV Predictor. In: JCT-VC 5th Meeting, JCTVC-E059, Geneva (2011)
7. Lim, S.C., Kim, H.Y., Kim, J., Lee, J., Choi, J.S.: Dynamic Range Restriction of Temporal Motion Vector. In: JCT-VC 5th Meeting, JCTVC-E142, Geneva (2011)
8. Su, Y., Segall, A.: CE9: Reduced Resolution Storage of Motion Vector Data. In: JCT-VC 4th Meeting, JCTVC-D072, Daegu, Korea (2011)
9. Choi, J., Kim, J.: Motion Vector Memory Reduction Scheme for Scalable Motion Estimation. *Optical Engineering* 48(9) (2009)
10. Fujibayashi, K., Bossen, F.: CE9 3.2d Modified Motion Vector Compression Method. In: JCT-VC 5th Meeting, JCTVC-E231, Geneva (2011)
11. Bossen, F.: Common Test Conditions and Software Reference Configurations. In: JCT-VC 4th Meeting, JCTVC-D600, Daegu, Korea (2011)
12. Meeting Report of The Fifth Meeting of The Joint Collaborative Team on Video Coding. In: JCT-VC 5th Meeting, Geneva (2011)
13. Wang, R., Li, J., Huang, C.: Motion Compensation Memory Access Optimization Strategies For H.264/AVC Decoder. In: *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 97–100 (2005)

A Local Variance-Based Bilateral Filtering for Artifact-Free Detail- and Edge-Preserving Smoothing

Cuong Cao Pham, Synh Viet Uyen Ha, and Jae Wook Jeon*

School of Information and Communication Engineering, Sungkyunkwan University,
Suwon, South Korea

cuongpc@skku.edu, synhha@ece.skku.edu, jwjeon@yurim.skku.ac.kr
<http://micro.skku.ac.kr>

Abstract. Edge-preserving smoothing has recently emerged as a crucial technique for a variety of computer vision and image processing applications. The idea is to smooth the small scale variations, while preserving edges and fine details in the image, without causing halo artifacts in detail enhancement. In this paper, we propose a modified bilateral filter model that has better behavior near edges and details than the standard model does. The edge-stopping function takes into account the intensity difference and the local variance of the filter windows, which provides insightful information about the local pixel distribution. We demonstrate the existing detail-related bilateral-based applications can achieve better results by simply switching from the standard model to our proposed model. In particular, we applied our method to detail-preserving image denoising and detail enhancement.

Keywords: Edge-preserving smoothing, bilateral filter, detail enhancement, image denoising.

1 Introduction

Edge-preserving smoothing plays a crucial role in a variety of applications in computer vision and image processing. The filtering process will smooth the noise, while preserving edges and fine details of the image. Both the smoothing result or base layer and residual detail layer can be used in different ways, depending on each application. For instance, the details can be boosted in applications, such as detail enhancement, while in image de-noising they will be discarded.

* This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2011-(C1090-1121-0008)), and by Priority Research Centers Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0018397).

State of the art techniques that enable this kind of filter include anisotropic diffusion (AD) [19], bilateral filter (BLF) [23], weighted least squares filter (WLS) [12] and guided image filter (GIF) [14]. Each has its advantages and disadvantages, so we need to consider the balance between the quality of the results and the computational cost. By taking this consideration, GIF and BLF would be the most suitable choices. Especially, BLF has become the *de facto* standard for computer vision and image processing applications due to its simplicity, flexibility and appealing characteristics [18].

The main challenges of BLF are its time consumption and production of reversed gradient artifacts near edges in detail enhancement technique. While the time taken can be reduced using fast approximated techniques [7,10,17,20], causing reversed gradient artifacts in detail enhancement and filtering meaningful fine details in image de-noising are unresolved issues.

In this paper, we propose a modified BLF model that has better behavior near edges than the standard model does. We realize that the halo artifacts, that occur in detail enhancement, come from the weak discrimination between pixel pairs in the local window. In general, the simple intensity difference of the standard model is insufficient to determine which pixels should be smoothed and which detail- and edge-pixels should be preserved. The proposed method takes into account the intensity difference as does the standard model, and the local variances of the image, which provide insightful information about local pixel distribution. Experiments show bilateral-based detail enhancement and detail-preserving image denoising can achieve better results by simply switching from the existing standard model to our improved model.

The remainder of this paper is organized as follows. Section 2 discussed various existing edge-preserving smoothing techniques. Section 3 presents the proposed method with the use of local variance. Section 4 presents experimental results to compare our method to methods from the literature. Finally, this paper is drawn to a conclusion and future work outlined in Section 5.

2 Related Work

Several state of the art edge-preserving smoothing techniques have been developed and applied in a variety of computer vision and image processing applications in the last decade. The ideal edge-preserving filter must smooth the small scale variations or noise of the image and preserve edges and fine details without causing any halo artifacts near edges in detail enhancement. In particular, it must neither blur nor sharpen the edges.

Such a filter was first used for scale-space image segmentation and edge detection known as anisotropic diffusion (AD) [19]. Generally, AD strongly smooths the low gradient regions, while preventing averaging across the strong edges where the gradient magnitude is large. This is the reason why AD can smooth noise while preserving edge structures. However, it has three major drawbacks. First, the meaningful fine details of an image are unexpectedly removed due to its over-smooth characteristic. Second, the edges are over-sharpened in the fi-

nal cumulative result. Finally, the non-linear iterative process of AD is slow to converge.

Although there are some improved anisotropic diffusion models, such as robust anisotropic diffusion [3], variance-based [6], LCIS [24], these limitations cannot be overcome thoroughly. BLF is increasingly widely used due to its appealing characteristics. It was originally developed in [1,21], and then popularized in [23] with the name bilateral filter. It is a non-iterative, non-linear weighted averaging filter that smooths noise, whilst preserving edges. A weighted mean of neighbor pixels is calculated to produce each output pixel in the filtered result. The weight is computed by the product of the spatial and intensity range domain. Let I_p be the intensity value at pixel p ; then, BLF is defined by:

$$BLF(I)_p = \frac{\sum_q g_{\sigma_s}(p, q) f_{\sigma_r}(I_p, I_q) I_q}{\sum_q g_{\sigma_s}(p, q) f_{\sigma_r}(I_p, I_q)} \quad (1)$$

where kernel functions g_{σ_s} and f_{σ_r} are spatial and range domain, respectively. Parameters σ_s and σ_r determine the weights in each domain, represented by a Gaussian function:

$$g_{\sigma_s}(p, q) = \exp(-\|p - q\|^2 / 2\sigma_s^2) \quad (2)$$

$$f_{\sigma_r}(I_p, I_q) = \exp(-|I_p - I_q|^2 / 2\sigma_r^2) \quad (3)$$

BLF can perform the smoothing process in a single iteration and overcome the over-sharpen edges characteristic of anisotropic diffusion. The weights assigned to neighbor pixels in spatial domain depend on their locations, while intensity values determine the weights in range domain. Specifically, the spatial domain reduces the weight of distant pixels, and the range domain reduces the weight of pixels that are different from the center pixel in terms of intensity value. The degree of smoothing is controlled by adjusting the value of σ_r . BLF becomes equivalent to the Gaussian low-pass filter when σ_r increases. Hence, this value must be sufficiently small to avoid filtering fine features.

The main disadvantages of BLF are its time consumption and production of gradient reversal artifacts near edges in detail enhancement. Several fast approximated techniques have been investigated to reduce the time-taken [7,10,17,20]. The main concepts of these acceleration schemes can be found in [18]. The halo artifacts that occur in detail enhancement can be controlled using a post-processing algorithm proposed in [15]. Our improved model greatly reduces these artifacts, without using such a post-processing step. Moreover, the detail-preserving characteristic of our model is another advantage compared to the standard model.

Some improvements of BLF that applied adaptive parameters instead of fixed parameters have been developed. Wong [26] included the local phase coherence into both spatial and range domain of BLF to improve the noise reducing and detail-preserving ability. However, the gradient reversal artifacts and detail enhancement have not been evaluated and discussed. While the adaptive BLF

developed in [27] tends to enhance the sharpness of the noisy blurred image, not to boost the detail as our stated problem.

To address the second problem of BLF, Farbmán et al. [12] also introduced the weighted least squares (WLS) filter that better preserves the edges than BLF does and is able to perform halo-free image decomposition of an input image. It can produce a high quality smoothed and enhanced result, without suffering obvious artifacts, but requires solving a large sparse linear system. This leads us to another challenge.

Subr et al. [22] favored using the local extrema-based filter (LEF) instead of WLS filter. The output image is computed by averaging min/max envelopes, which are interpolated from min/max extrema using the edge-preserving interpolation scheme [16]. However, this method tends to remove details even in high contrast textured regions. Furthermore, it causes incomplete smoothing of the oscillation in a single iteration, and requires solving k number of large sparse linear system ($k \geq 2$).

In a notable recent work, He et al. [14] made use of the local linear transform model to propose guided image filtering (GIF). Although GIF can reduce the gradient reversal artifacts of BLF, and can be implemented using a fast algorithm, the details are not well preserved, and the contrast of its resultant output is reduced, as in the case of BLF.

In summary, we need to trade off the result quality and computational cost. Among existing methods, the bilateral-based filter is still widely used and would have become the *de facto* standard for computer vision applications [18]. The next section presents our improved BLF model with the use of local variance. Experiments show our method can achieve better results than conventional BLF in detail enhancement and image de-noising.

3 Proposed Bilateral Filter Model

In this section, we present a modified BLF model that better preserves edges and fine details compared to the standard model. The edge-stopping function will take into account the intensity difference and the local variance of the filter windows. First, we briefly analyze the variance information of the local windows, and then apply it to the proposed model. The related analysis can be found in [6] for degraded background images, while our analysis strongly focuses on the windows nears edges.

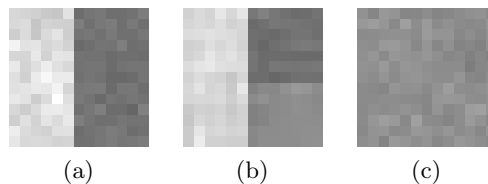


Fig. 1. Local windows of size 13×13 of an input image

Table 1. Comparison of our method and existing edge-preserving smoothing techniques

Filtering	Smoothed	Preserved	Assumption
AD	Low gradient region	Large gradient region	Gradient magnitude at edges is large
BLF	Low gradient region	Large intensity difference	Noise is low contrast
GIF	Low variance region	High variance region	Noise is low contrast
WLS	Low gradient region	Large gradient region	Gradient magnitude at edges is large
LEF	Oscillations between min and max local extrema	High variance of neighboring local extrema	Detail can be low or high contrast
Our	Low gradient, low variance region	Large intensity difference, high variance region	Noise is low contrast

3.1 Local Variance Analysis

For the enlarged windows of an input image as shown in Fig. 1. Fig. 1(a) and 1(b) contains some step edges, while Fig. 1(c) only contains small scale oscillations. The objective of edge-preserving smoothing for such these images is to smooth the small scale oscillations and prevent their inter-region edges leaking together. The variances of these windows are 11.14, 8.44 and 0.18, respectively. It has long been known that the local variance is calculated by:

$$\sigma_p^2 = \frac{1}{|w|} \sum_{q \in w_p} (I_q - \mu_p)^2 \quad (4)$$

where μ_p is the mean of I in the local window w_p , and $|w|$ is the number of pixels in w_p . As discussed in [6], it should be normalized to make the range of variance to be compatible with the range of gradient, as in equation (5):

$$\sigma_{p,N}^2 = 1 + \frac{\sigma_p^2 - \min \sigma_p^2}{\max \sigma_p^2 - \min \sigma_p^2} .254 \quad (5)$$

where $\min \sigma_p^2$ and $\max \sigma_p^2$ are the minimum and maximum local variance of the entire image. The normalized variances of these above windows are 234.79, 178.13 and 4.61, respectively.

We can observe that the variances of the local windows that contain inter-region edges are larger than that in homogeneous regions. In order to preserve detail- or edge-pixels that located near the center of these windows, their neighbor pixels should have small weight in the filtering process. For instance, while computing the output of center pixels in Fig. 1(a) and 1(b), the small weights should be assigned to their neighbors. Meanwhile, a low-pass filter is applicable to homogeneous region, as Fig. 1(c).

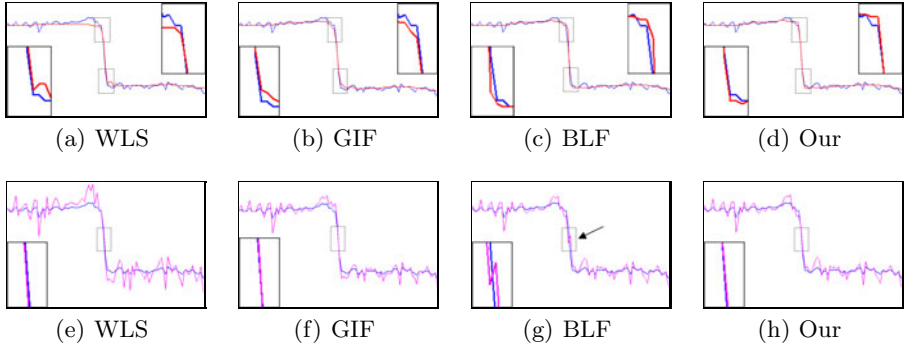


Fig. 2. Filter outputs (top row) and detail enhancement (bottom row) in 1-D of our method and existing methods. Parameters are configured as followed: WLS ($\lambda = 0.5$, $\alpha = 1.2$), GIF ($r = 6$, $\epsilon = 0.05^2$), BLF ($\sigma_s = 6$, $\sigma_r = 0.1$) and Our ($\sigma_s = 6$, $\sigma_{vr} = 0.01$)

3.2 Proposed BLF Model

Capture the notion of local variance and its insightful information about the local pixel distribution, we proposed the modified BLF model that takes into account both intensity difference and local variance. We tend to reduce the weights of local neighbors in the smoothing process of high-variance detail- or edge-pixels. The proposed BLF model with the use of local variance is given as follows:

$$VBF(I)_p = \frac{\sum_q g_{\sigma_s}(p, q) f'_{\sigma_{vr}}(I_p, I_q, \sigma_{p,N}^2) I_q}{\sum_q g_{\sigma_s}(p, q) f'_{\sigma_{vr}}(I_p, I_q, \sigma_{p,N}^2)} \quad (6)$$

where g_{σ_s} is spatial domain defined in (2), and $f'_{\sigma_{vr}}$ is our modified edge-stopping function,

$$f'_{\sigma_{vr}}(I_p, I_q, \sigma_{p,N}^2) = \exp\left(-\frac{1}{2} \left(\frac{|I_p - I_q| \cdot \sigma_{p,N}^2}{\sigma_{vr}}\right)^2\right) \quad (7)$$

where σ_{vr} plays the role of the smoothing parameter, the larger the value of σ_{vr} is, the smoother the filter image will be. However, it must be much smaller than the standard deviation σ_r used in the standard model in the normalized $[0; 1]$ intensity range value. With this proposed edge-stopping function, the small weights will be assigned to the neighbors of high-variance detail- and edge-pixels, this effect prevents two inter-region edges leaking together as the ideal filter. Meanwhile, the variances are close to one in the homogeneous regions; the proposed model becomes equivalent to the standard BLF, and noise and small scale oscillations will be smoothed. Table 1 shows the overview characteristics of our method in the context of existing techniques.

3.3 Edge-Preserving Characteristic

Fig. 2 visually compare our method and existing methods. Given an 1-D input signal (blue) plotted from a scan-line of an image, the filter outputs (red) are

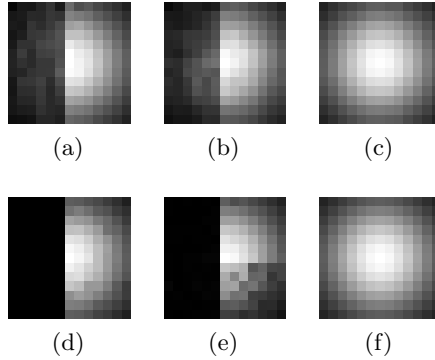


Fig. 3. Filter kernels of bilateral filter (top row) and our method (bottom row) corresponding to Fig. 1(a), 1(b), and 1(c)

shown in the top row, and the enhanced outputs (magenta) are shown in the bottom row. We can easily see that the reversed gradient artifacts only in the enhanced output of BLF (Fig. 2(g)), because the edge-pixel and its neighbors have leaked together in the smoothed result (base layer) (Fig. 2(c)). In contrast, our method with the participation of local variance information does not suffer such artifacts, as shown in Fig. 2(h). The WLS filter and GIF yield high quality in both smoothed (Fig. 2(a) and 2(b)) and enhanced output (Fig. 2(e) and 2(f)), as they have been proved previously [12,14].

Fig. 3 shows the filter kernel visualization of BLF and our method corresponding to the enlarged windows of Fig. 1. Fig. 3(a) and 3(b) show the weights assigned to the pixels on the non-same region to the center pixel are still large, so the edges will not be consistently preserved for BLF. Conversely, the weights assigned to these pixels in our method are nearly zero (black), as shown in Fig. 3(d) and 3(e). This better preserves the edge-pixels than the standard BLF does. Both filters have equivalent behavior for the low-gradient, low-variance regions, the neighbors are averaged together. As we can see in the right-most column, both filter kernels are comparable to a low-pass filter.

3.4 Implementation

The proposed model naturally works well with a gray-scale image. In order to process with color image, both intensity difference and local variance are calculated by Euclidean distance in CIELab color space. We can achieve better results by processing in CIELab space, rather than RGB, because the channel-wise CIELab space is more uniform than RGB [23] is. The algorithm is directly implemented using CUDA in this work to achieve the speed-up. The GPU implementation takes about 90ms to process 1-megapixel gray-scale image, and 120ms for color image. The measurement was performed in a PC with an AMD Athlon 64 X2 Dual Core Processor 3800+ 2.00 Ghz and NVIDIA GeForce GT 240.

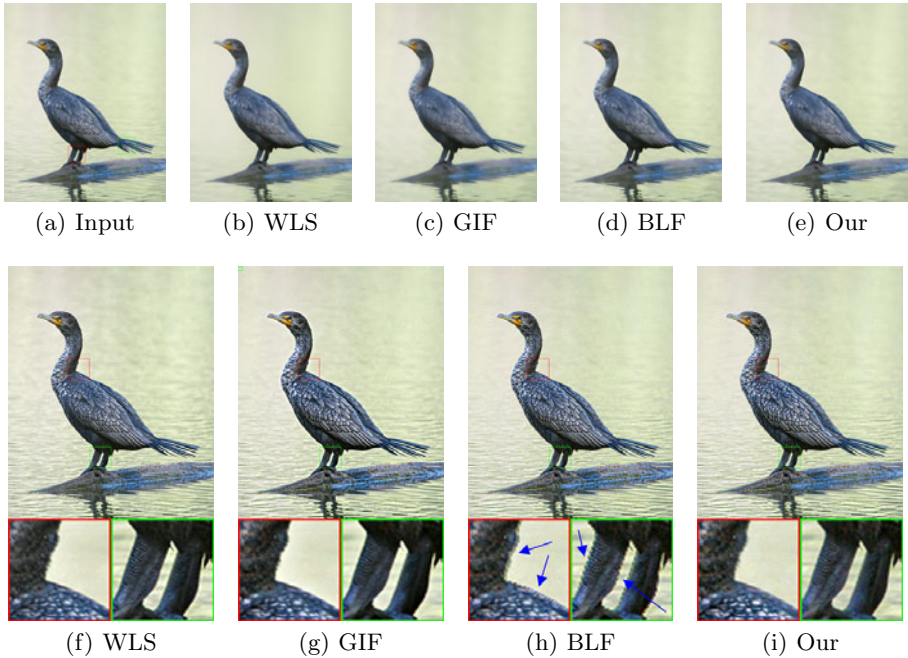


Fig. 4. Filter outputs (top row) and detail increased by $3\times$ (bottom row) of our method and existing methods for color image. Parameters are configured as follows: WLS: $\lambda = 0.2$, $\alpha = 1.2$; GIF: $r = 6$, $\epsilon = 0.01$; BLF: $\sigma_s = 6$, $\sigma_r = 0.1$; Our: $\sigma_s = 6$, $\sigma_{vr} = 0.01$. The input image was taken from [15].

4 Experimental Results

4.1 Detail Enhancement

Our proposed method focuses on the gradient reversal artifacts occurring in the detail enhancement technique, where the details are first boosted, and then recombined with the base layer to produce an enhanced output. Fig. 4 shows the enhanced results rendered by our method and existing methods. When the artifacts occur near the edges in the BLF result (Fig. 4(h)), our method (Fig. 4(i)) greatly reduces these artifacts and can be visually comparable to the enhanced output produced from WLS filter (Fig. 4(f)) and GIF (Fig. 4(g)). The same analysis can be applied to the enhanced outputs of gray-scale image as shown in Fig. 5. The reversed gradient artifacts only occur in the BLF result, while the enhanced output produced from our method is comparable to that produced from WLS and GIF.

4.2 Detail-Preserving Image Denoising

Denoising is the first and original purpose of image filtering. Especially, smoothing noise, while preserving both edges and fine details in the image simultaneously, is

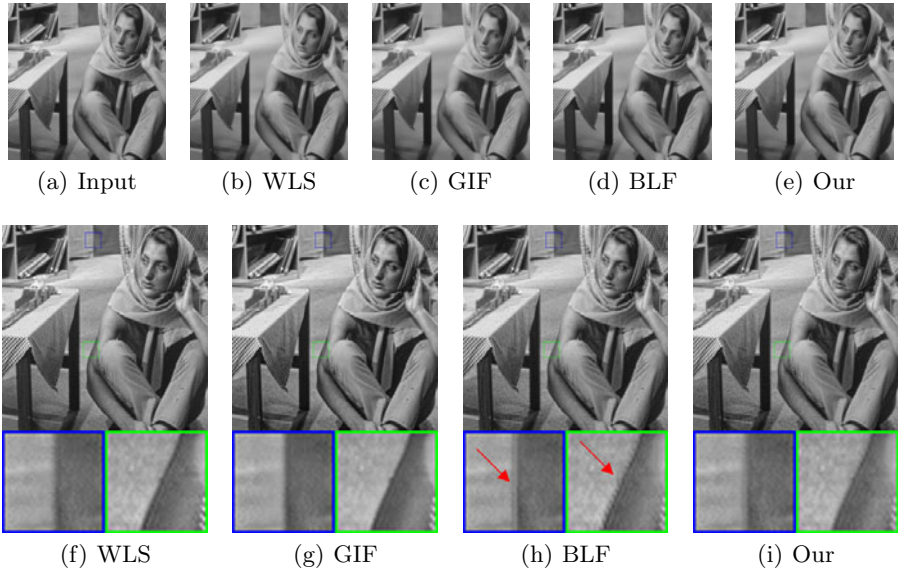


Fig. 5. Filter outputs (top row) and detail increased by $3\times$ (bottom row) of our method and existing methods for gray-scale image. Parameters are configured as follows: WLS: $\lambda = 0.1$, $\alpha = 1.2$; GIF: $r = 6$, $\epsilon = 0.05^2$; BLF $\sigma_s = 6$, $\sigma_r = 0.05$; Our: $\sigma_s = 6$, $\sigma_{vr} = 0.05^2$.

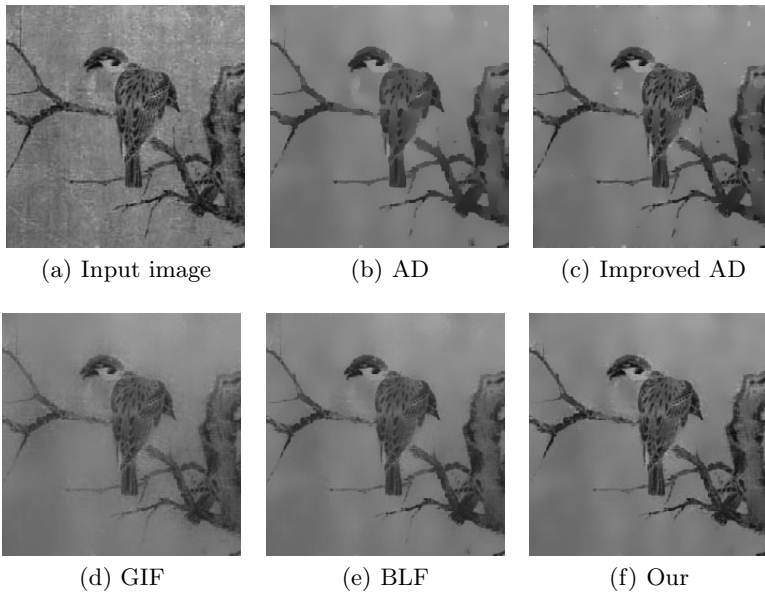


Fig. 6. Comparison of detail-preserving image denoising. Parameters are configured as follows: (b) AD: $K = 3$; (c) Improved AD: $K_0 = 12$; (d) GIF: $r = 10$, $\epsilon = 0.01$; (e) BLF: $\sigma_s = 10$, $\sigma_r = 0.1$; (f) Our method: $\sigma_s = 10$, $\sigma_{vr} = 0.12^2$. The input image was taken from [6].

an important and challenging task [6]. In terms of restoration using filtering without deploying any noise detection operator, we compare our proposed method with various edge-preserving smoothing techniques including AD [19], improved AD [6], GIF [14] and BLF [23]. The test image is the degraded bird image taken from [6]. While the AD and the improved AD are set to be iterated over 100 times, the smoothing parameters of remaining methods are chosen carefully to achieve tolerable results. The restored image in AD (Fig. 6(b)) has lost most of the details of the bird and the edges are over-sharpened. Though the improved anisotropic diffusion preserves both edges and fine details in the restored image (Fig. 6(c)), it cannot avoid sharpening edges, the same as in anisotropic diffusion. Both GIF and BLF effectively smooth noise in the homogeneous regions, however, the details have been blurred and contrast is reduced, as shown in Fig 6(d) and 6(e), respectively. Our method preserves edges and fine details and avoids sharpening edges, as shown in Fig. 6(f) at the expense of small noise remained near edges. Moreover, the contrast is closer to that of the input image.

5 Conclusion

In this paper, we presented an improved BLF model for artifact-free detail- and edge-preserving smoothing. The improved model takes into account not only the intensity difference but also the local variance information of the filter windows. It better preserves edges and fine details compared to the standard model. The gradient reversal artifacts have been greatly reduced without using a post-processing step in detail enhancement, while the meaningful fine details are also preserved in the image denoising technique.

Experimental results showed these detail-related bilateral-based applications can achieve better results by simply switching from the standard model to our proposed model. Its ability to apply to a variety of applications, such as multi-scale image decomposition, HDR compression, and flash/no-flash imaging, is limited compared to WLS filter and guided image filter. Further improvement, quantitative evaluations and its relation to non-local filters should be investigated. We leave these challenges for our future works.

References

1. Aurich, V., Weule, J.: Non-linear gaussian filters performing edge preserving diffusion. In: Proceedings of the DAGM Symposium, pp. 538–545 (1995)
2. Barash, D.: A Fundamental Relationship Between Bilateral Filtering, Adaptive Smoothing, and the Nonlinear Diffusion Equation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(6), 844–847 (2002)
3. Black, M.J., Sapiro, G., Marimont, D.H., Heeger, D.: Robust anisotropic diffusion. *IEEE Transactions on Image Processing* 7(3), 421–432 (1998)
4. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: *IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 60–65 (2005)

5. Buades, A., Coll, B., Morel, J.M.: The staircasing effect in neighborhood filters and its solution. *IEEE Transactions on Image Processing* 15(6), 1499–1505 (2006)
6. Chao, S.M., Tsai, D.M., Chiu, W.Y., Li, W.C.: Anisotropic diffusion-based detail-preserving smoothing for image restoration. In: *IEEE Intl. Conf. on Image Processing (ICIP)*, pp. 4145–4149 (2010)
7. Chen, J., Paris, S., Durand, F.: Real-time edge-aware image processing with the bilateral grid. *ACM Transactions on Graphics* 26(3) (2007)
8. Choudhury, P., Tumblin, J.: The trilateral filter for high contrast images and meshes. In: *Proceedings of the Eurographics Symposium on Rendering*, pp. 186–196 (2003)
9. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: BM3D Image Denoising with Shape-Adaptive Principal Component Analysis. In: *Workshop on Signal Processing with Adaptive Sparse Structured Representation (SPARS)* (2009)
10. Durand, F., Dorsey, J.: Fast Bilateral Filtering for the Display of High-Dynamic-Range Images. *ACM Transactions on Graphics* 21(3), 257–266 (2002)
11. Elad, M.: On the bilateral filter and ways to improve it. *IEEE Transactions on Image Processing* 11(10), 1141–1151 (2002)
12. Farbman, Z., Fattal, R., Lischinski, D., Szeliski, R.: Edge-Preserving Decompositions for Multi-Scale Tone and Detail Manipulation. *ACM Transactions on Graphics* 27(3) (2008)
13. Gupta, M.D., Xiao, J.: Bi-affinity Filter: A Bilateral Type Filter for Color Images. In: *ECCV Workshop on Color and Reflectance in Computer Vision, CRICV* (2010)
14. He, K., Sun, J., Tang, X.: Guided Image Filtering. In: *Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311*, pp. 1–14. Springer, Heidelberg (2010)
15. Kass, M., Solomon, J.: Smoothed Local Histogram Filters. *ACM Transactions on Graphics* 29(4) (2010)
16. Levin, A., Lischinski, D., Weiss, Y.: Colorization using Optimization. *ACM Transactions on Graphics* 23(3), 689–694 (2004)
17. Paris, S., Durand, F.: A Fast Approximation of the Bilateral Filter using a Signal Processing Approach. *International Journal of Computer Vision* 81(1), 24–52 (2009)
18. Paris, S., Kornprobst, P., Tumblin, J., Durand, F.: Bilateral Filtering: Theory and Applications. In: *Foundations and Trends in Computer Graphics and Vision* (2009)
19. Penora, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12(7), 629–639 (1990)
20. Pham, T.Q., Van Vliet, L.J.: Separable bilateral filtering for fast video preprocessing. In: *Proceedings of the IEEE Intl. Conf. on Multimedia and Expo* (2005)
21. Smith, S.M., Brady, J.M.: SUSAN - A new approach to low level image processing. *International Journal of Computer Vision* 23(1), 45–78 (1997)
22. Subr, K., Soler, C., Durand, F.: Edge-preserving Multiscale Image Decomposition based on Local Extrema. *ACM Transactions on Graphics* 28(5) (2009)
23. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: *Proceedings of the IEEE Intl. Conf. on Computer Vision (ICCV)*, pp. 839–846 (1998)
24. Tumblin, J., Turk, G.: LCIS: A boundary hierarchy for detail-preserving contrast reduction. In: *ACM SIGGRAPH 1999*, pp. 83–90 (1999)
25. Weickert, J.: *Anisotropic Diffusion in Image Processing*, Stuttgart, Germany (1998)
26. Wong, A.: Adaptive bilateral filtering of image signals using local phase characteristics. *Signal Processing* 88(6), 1615–1619 (2008)
27. Zhang, B., Allebach, J.P.: Adaptive Bilateral Filter for Sharpness Enhancement and Noise Removal. *IEEE Transactions on Image Processing* 17(5), 664–678 (2008)

Iterative Gradient-Driven Patch-Based Inpainting

Sarawut Tae-o-sot¹ and Akinori Nishihara²

¹ The Department of Communications and Integrated Systems,
Tokyo Institute of Technology, Tokyo, Japan
`sarawutt@nh.cradle.titech.ac.jp`

² The Center for Research and Development of Educational Technology,
Tokyo Institute of Technology, Tokyo, Japan
`aki@cradle.titech.ac.jp`

Abstract. A novel exemplar-based image inpainting is proposed in this paper. This method is based on iterative approach which provides better result than greedy one. The problem of inconsistent results caused by raster scanning on target patch selection in iterative approach is focused in this paper. The proposed gradient-driven ordering is used to select target patch instead of traditionally predefined ordering. Due to the information-driven nature, this new approach is image's rotation invariant which means the same result is provided by different rotation of the same damaged image. Moreover, a random search approach is redesigned to be more reasonable and suitable for our novel gradient-driven ordering. The proposed method provides the best inpainting result among several well-known exemplar-based inpainting techniques including both greedy and iterative approach.

Keywords: image completion, image inpainting, exemplar-based, patchmatch.

1 Introduction

Image inpainting is the research area in the field of image processing whose goal is to remove some objects or restore damaged regions in a way that observers cannot notice the flaw. There are many applications of image inpainting such as photo editing, video editing, image compression and image transmission. Generally image inpainting techniques can be categorized into two approaches, Diffusion-based and Exemplar-based approach.

Diffusion-based approach is the fundamental approach in which information diffuses from known region into missing region. The problem is usually modelled by Partial Differential Equation (PDE), so sometimes it is called a PDE-based approach. Bertalmio *et al.* [2] reconstructed missing regions by diffusing known region along isophote direction, the direction of equal intensity value, into the missing region by a heat flow model. Chan *et al.* [4] introduced Total Variance (TV) framework for inpainting problem, then curvature-driven diffusion (CDD)

[5] which fixes connectivity problem in TV model. Diffusion-based approach works well for non-texture image, in which the missing region must be small and thinner than the surrounding object. In the case that the missing region is large or containing texture, this approach gives a blurry result.

Exemplar-based approach is originated from the Exemplar-based texture synthesis in [7] which synthesizes texture by copying the best match patch from the known region. However, directly applying exemplar-based texture synthesis to image inpainting problem may not provide satisfactory result. This is because, there are both structures and textures in natural images. Bertalmio [3] proposed to decompose the image into structural and textural images, then apply diffusion-based inpainting to the structural image and texture synthesis to the textural image separately. The result of combining restored structural and textural image is better than restoration by only diffusion-based inpainting or texture synthesis alone. However, that technique still cannot recover the large missing region. Criminisi *et al.* [6] introduced patch priority, which is defined by isophote direction and known pixels in target patch, for exemplar-based texture synthesis to determine the fill-in order. In that way, the structural information is recovered because the target patches which have high structural information are likely to be filled first. Kwok *et al.* [10] introduced DCT-based inpainting in which patch matching process is done in DCT domain. In that way, the error which is caused by noise is reduced by the noise reduction properties of DCT. However, new error is produced by the gradient-based filling process which roughly approximates the unknown region of the target patch before doing DCT. Patch shifting approach [13] was introduced for enhancing [6] and [10] performance. The main idea of the approach is to simultaneously reduce the unknown region and increase the known region of each target patch.

The major drawback of greedy approaches mentioned earlier is that filled region cannot be refilled. The error of the prior filled patches affect the next filled patch. This means that the error is accumulated. Due to this difficulty, iterative approach was considered. Komodakis *et al.* [9] modelled image inpainting problem by Markov Random Field and introduced priority belief propagation for solving. Graph cuts technique was introduced to optimize the problem in the work of Lee *et al.* [11]. Exemplar-based inpainting proposed by Wexler *et al.* [14] modelled inpainting as global optimization problem. Unlike [6], unknown region is filled iteratively until the solution converges. Generally, the iterative approach provides better results by sacrificing the computational time. However, patchmatch technique [1], which is a fast and efficient tool for dealing with exemplar-based problems, can reduce this huge complexity. A fast computational time is the result of random search strategy which compromises with the final result. For the best result of patchmatch inpainting, the structure of damaged area need to be manually specified. Unlike most of the greedy approaches, target patches are in the raster scan order which leads to inconsistency results. This means restored results of the same damaged image with different rotation may be different.

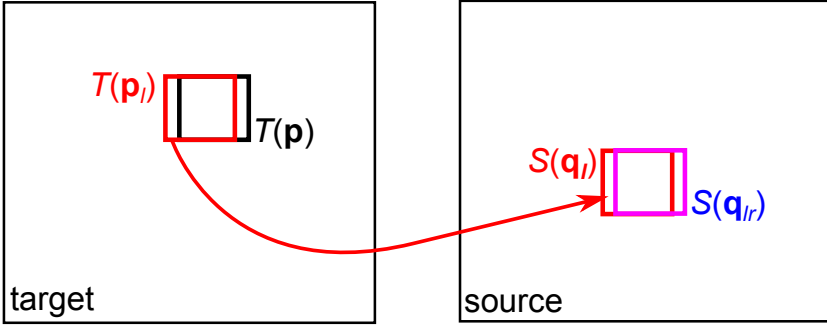


Fig. 1. Propagation step: candidate suggested from left neighbour of $T(\mathbf{p})$

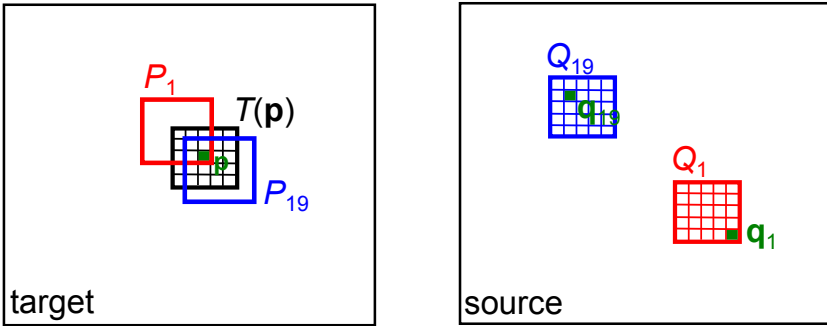


Fig. 2. Image reconstruction: Example of how corresponding pixel is defined

In this paper, iterative gradient-driven exemplar-based image inpainting, which is inspired by a fast but inconsistent patchmatch method, is proposed. Gradient-driven ordering is introduced in our proposed method. With this ordering, fill-in orders are arranged by the strength of gradient in target patch instead of traditional raster scanning. In other word, the patch with high structural information receives a priority to do patch matching first. Moreover, random search strategy is redesigned to be more efficient and able to support gradient-driven ordering scheme. By our novel framework the better and more consistent result can be achieved with almost the same speed as patchmatch inpainting.

This paper is organized as follows. In section 2, a new random search strategy is presented. In section 3, Gradient-driven ordering is introduced. Image reconstruction from nearest-neighbour field is explained in section 4. The experimental result of our framework and several inpainting techniques are presented and discussed in section 5. Finally, the conclusion of our work and the direction of our work are mentioned in section 6.

2 Random Search Strategy

Generally image inpainting is modeled as the problem of filling-in the missing region Ω , sometimes called target region, of the given image domain U by the information of the known region, sometimes called source region $U \setminus \Omega$. In exemplar-based view point, this is the problem of searching for the best match patch of all target patch T , where $T \in \Omega$, from the source region $U \setminus \Omega$. Exhaustive search was used in [14] and the results were excellent. Undoubtedly, computational complexity of this method is very high which is mathematically $O(mM_\Omega M)$, where m , M_Ω and M are numbers of pixels in patch, unknown region, Ω , and known region, Φ , respectively. The faster search method was introduced in [1]. The complexity of this method is around $O(mM_\Omega \log M)$ and can be optimized to $O(\sqrt{m}M_\Omega \log M)$. Cooperating with multi-resolution image processing patchmatch can be implemented in almost real-time.

Our search strategy is inspired by [1], however, many details are changed in order to make it compatible with our gradient-driven ordering scheme. Basically, the search strategy is the combination of 3 steps including initialization, propagation and random search.

2.1 Initialization

In this stage, all the patches in unknown region are assigned its corresponding patch randomly. In case of multi-resolution processing, the corresponding patches are initialized by the upscaled version of the previous low resolution level.

2.2 Propagation

Target patch $T(\mathbf{p})$ will consider the 4 candidate patches from the suggestion of its 4-neighbour patches including left neighbour $T(\mathbf{p}_l)$, right neighbour $T(\mathbf{p}_r)$, upper neighbour $T(\mathbf{p}_u)$ and lower neighbour $T(\mathbf{p}_d)$. The candidate of each neighbour is suggested from a viewpoint of its corresponding patch. For example, considering the left neighbour $T(\mathbf{p}_l)$ in Fig. 1, if $S(\mathbf{q}_l)$ is the corresponding patch of $T(\mathbf{p}_l)$, it would be possible that the right neighbour of $S(\mathbf{q}_l)$ is the corresponding patch of $T(\mathbf{p})$. So, $S(\mathbf{q}_{lr})$, the blue patch, is chosen as the candidate. In the same manner, $T(\mathbf{p}_r)$, $T(\mathbf{p}_u)$ and $T(\mathbf{p}_d)$ will suggest $S(\mathbf{q}_{rl})$, $S(\mathbf{q}_{ud})$ and $S(\mathbf{q}_{du})$ as candidates. Finally, the new corresponding patch of $T(\mathbf{p})$ can be calculated from

$$T(\mathbf{p}) = \arg \min_S \{d(T(\mathbf{p}), S(\mathbf{q}_{lr})), d(T(\mathbf{p}), S(\mathbf{q}_{rl})), \\ d(T(\mathbf{p}), S(\mathbf{q}_{ud})), d(T(\mathbf{p}), S(\mathbf{q}_{du})), d(T(\mathbf{p}), S(\mathbf{q}))\}, \quad (1)$$

where $d(T, S)$ is Euclidean distance of patch T and S .

2.3 Random Search

This step randomly picks up the candidate from the searching region whose radius exponentially decreases. Unlike [1], the center of searching region may



Fig. 3. The experiment on restoring known image: (a) the original image, (b) the damaged image, (c) the result of Criminisi's method with PSNR = 35.42 dB, (d) the result of DCT method with PSNR = 36.17 dB, (e) the result of Wexler's method with PSNR = 37.28 dB, (f) the worst result of patchmatch inpainting with PSNR = 38.21 dB, (g) the best result of patchmatch inpainting with PSNR = 39.15 dB, (h) the result of our proposed method with PSNR = 39.35 dB

not be fixed to the center of the corresponding patch from propagation step \mathbf{q}_0 . At every decrease in searching radius a new corresponding patch is determined by the previous corresponding patch $S(\mathbf{q}_{i-1})$ and the random candidate $S(\mathbf{q}_i)$. If the corresponding patch changes, the center for randomly selecting the candidate patch will also change. This procedure would be repeated until the search radius is less than 1. A complete strategy of this step can be described as follows,

initialization the center of searching region is set to the center of corresponding patch \mathbf{q}_0 .

iteration

-new candidate is chosen by

$$\mathbf{q}_i = \mathbf{q}_{i-1} + w\alpha^i \mathbf{R}_i, \quad (2)$$

where w is the maximum search radius, α is reduction ratio of search radius and \mathbf{R}_i is a uniform random vector in $[-1, 1] \times [-1, 1]$. Normally, α is set to 0.5.

-the new corresponding patch is updated by

$$S(\mathbf{q}_i) = \arg \min_S \{d(T(\mathbf{p}), S(\mathbf{q}_{i-1})), d(T(\mathbf{p}), S(\mathbf{q}_i))\} \quad (3)$$

With this novel strategy, the corresponding patch seems to converge faster than the traditional strategy. A random candidate patch in each iteration always locates around the corresponding patch of each iteration. In contrast, the random candidates of [1] are densely around candidate patch of propagation step $S(\mathbf{q}_0)$. This condition may impede the convergence.

3 Gradient-Driven Ordering

On exemplar-based image inpainting, a predefined filling order, such as raster scan or onion peel, usually leads to unsatisfactory result as discussed in [6]. Moreover, this also leads to an inconsistent result. Not only greedy approach but also iterative approach, which is affected by predefined filling order. As shown in Fig. 3(f) and (g) and 4, with different raster scan directions, different results are achieved. In other word, the reconstruction results of same damaged image but different rotations, e.g., the damaged image and its 90° rotation, are probably different. And we do not know which rotation would provide the best result.

In this paper, structural information of patch is involved in ordering process. Magnitude of gradient is used as a measure of structure. Generally, on human point of view, a good restored image should contain no discontinuous structure [8]. To maintain this criterion, the patch with high magnitude of gradient should have a high priority to be selected as a target patch. Anyway, reliability of target patch must also be considered. So, the patches on the boundary of unknown region, which should be more reliable than ones inside the unknown region, must be considered first.

The concept of Gradient-driven ordering is to select the patch on the boundary of unknown region with the highest magnitude of gradient first as target patch. Then searching for its corresponding patch by random search strategy in section 2. After the target patch get its correspondence, the new boundary is defined by ignoring the target patch. And new target patch is selected again from this new boundary. This procedure is repeated until all patches in unknown region have

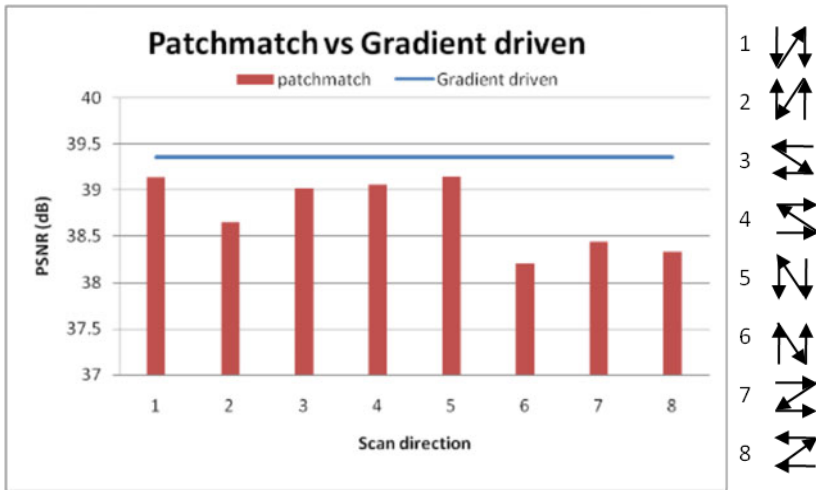


Fig. 4. PSNR of restoration of 3(b) by all scanning direction of patchmatch vs. our proposed method

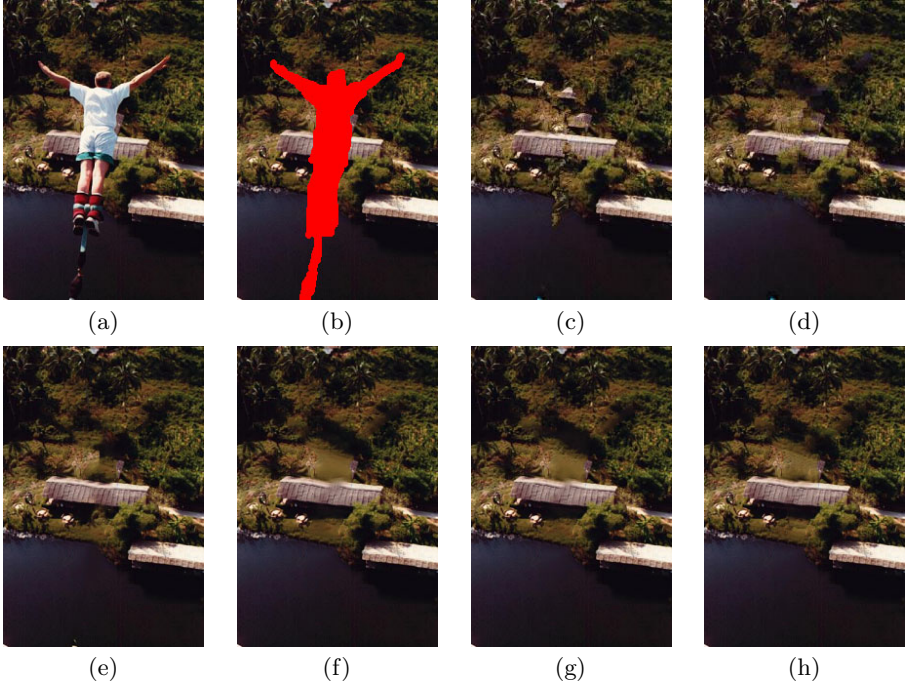


Fig. 5. The experiment on removing undesired object: (a) the original image of bungee jumping man, (b) removed region, (c) the result of Criminisi’s method, (d) the result of DCT method, (e) the result of Wexler’s method, (f) result of patchmatch inpainting with vertical raster scan from top-left to bottom-right corner, (g) result of patchmatch inpainting with horizontal raster scan from top-right to bottom-left corner, (h) the result of our proposed method

their corresponding patch. The whole process of gradient-driven ordering can be described as follows,

- initialization
 - Set all known pixels as “freeze”.
 - Set all unknown pixels as “active”.
- iteration
 - Step 1 Calculate magnitude of gradient on the boundary of “active” region.
 - Step 2 Select patch with maximum magnitude of gradient as target patch $T(\mathbf{p}_{max})$.
 - Step 3 Search for the corresponding patch of $T(\mathbf{p}_{max})$.
 - Step 4 Set \mathbf{p}_{max} as “freeze”.
 - Step 5 If there is still “active” pixel left, go to step 1.

4 Image Reconstruction

After finishing random search procedure for all $T(\mathbf{p}_i)$ where $\forall \mathbf{p}_i \in \Omega$, each $T(\mathbf{p}_i)$ then has its corresponding patch $S(\mathbf{q}_i)$. This section shows how to reconstruct the unknown region from all corresponding patches $S(\mathbf{q}_i)$. Denote P_1, P_2, \dots, P_M as all the patches that contain \mathbf{p} of $T(\mathbf{p})$. The number of patch M is equal to the size of the patch. Q_1, Q_2, \dots, Q_M denote the corresponding patches of P_1, P_2, \dots, P_M . And $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_M$ denote the corresponding pixel of \mathbf{p} suggested by Q_1, Q_2, \dots, Q_M respectively. For example, the patch Q_1 would suggest \mathbf{q}_1 for \mathbf{p} and Q_{19} would suggest \mathbf{q}_{19} for \mathbf{p} as shown in Fig. 2. In general, the best solution should compromise all the suggestion. So, the restored intensity $\hat{I}(\mathbf{p})$ should minimize the error from all suggestions. This error $e(I(\mathbf{p}))$ can be defined as follows,

$$e(I(\mathbf{p})) = \frac{1}{M} \sum_{i=1}^M (I(\mathbf{q}_i) - I(\mathbf{p}))^2 \quad (4)$$

After differentiating $e(I(\mathbf{p}))$ with respect to $I(\mathbf{p})$ and equating to zero, $I(\mathbf{p})$ can be obtained from

$$\hat{I}(\mathbf{p}) = \frac{1}{M} \sum_{i=1}^M I(\mathbf{q}_i). \quad (5)$$

This is the update equation for pixel \mathbf{p} . To reconstruct the unknown region all pixel inside the region need to be updated by this equation. Actually, $\hat{I}(\mathbf{p})$ is an average of all the intensity of the candidate pixel $I(\mathbf{q}_i)$. Equation (5) is just the simple version of update equation in [12] where completeness term is ignored.

5 Experimental Results and Discussion

In our experiment, beside showing the restored image, numerical evaluation is also used. In the case where the original image is available, peak signal to noise ratio (PSNR) is used to evaluate the result. Multi-resolution processing is applied to all the implementation of our proposed method.

5.1 Uniqueness of the Proposed Method and Inconsistency of Iterative Approach

In this experiment, the effect of raster scanning direction is investigated. All raster scanning directions have been implemented. In Fig. 3(f) and (g), the result of patchmatch inpainting with 7×7 patch size but different scanning direction are shown. Fig. 3(f), whose PSNR is 38.21 dB, is the worst result of all scanning direction. This image is achieved by scanning on vertical direction from bottom-left to top-right of the image. Fig. 3(g), whose PSNR is 39.15 dB, is the best result of all scanning direction. This image is achieved by scanning on vertical direction from top-right to bottom-left of the image. Obviously, these

two direction are just the opposite of each other. And for patchmatch algorithm, these are the same process with switching scan order. These results show that the result of patchmatch is very sensitive to scanning direction. With gradient-driven ordering on 7×7 patch size, the result, shown in Fig. 3(h), looks better than 3(f) and (g). And PSNR of Fig. 3(h), which is 39.35dB, is also higher than 3(f) and (g). According to Fig. 4, the proposed method outperforms patchmatch inpainting on all directions.

Another example is shown in Fig. 5(f)-(h). The result of our proposed method, shown in Fig. 5(h), obviously look better than the results of patchmatch inpainting, shown in Fig. (f) and (g). And, also, the results of patchmatch inpainting are not unique when the scanning direction is changed.

5.2 The Performance of the Proposed Method vs. Other Well-Known Inpainting Methods

In this experiment, the performance of our proposed method is compared with several well-known inpainting techniques except patchmatch inpainting which has already been discussed in the previous experiment. Firstly, the artificial damaged image shown Fig. 3(b), generated from the original image Fig. 3(a), is used for evaluating the performances. Because of the availability of original image, the performances can be evaluated in both visual and numerical aspect. The best result of Criminisi's method with 9×9 patch size is shown in Fig. 3(c). PSNR of the result is 35.42dB. Fig. 3(d) shows the best result of DCT method where PSNR is 36.17 dB. Although PSNR of DCT method is higher than Criminisi's method, it is obviously seen that the result of Criminisi's method has more natural look. So, just only numerical result may not measure the performance of inpainting. The result of Wexler's method is shown in Fig. 3(e) whose PSNR is 37.28. In this image, some discontinuity can be noticed as seen on the shoulder of the model. The result of our proposed method with 7×7 patch size is shown in Fig.3(h). Our proposed method achieved the highest PSNR of 39.35 dB. Visually, Fig. 3(h) seems the best reconstructed image. The shoulder of the model is restored perfectly by our method while other techniques fail.

In Fig. 5, the performance for object removing task is shown. The Bungee jumping man, which is masked as red region in Fig. 5(b), on the original image in Fig. 5(a) would be removed by various techniques. Fig. 5(c)-(e) and (f) show the best result by Criminisi's method, DCT method, Wexler's method and our proposed method respectively. Some discontinuity with strong false edge can be noticed on the result of Criminisi's method with 5×5 patch size in Fig. 5(c). The result of smooth edge can be noticed in DCT method with 9×9 patch size as shown in Fig. 5(d). In Fig. 5(e), Wexler's method with 9×9 patch size gives the result with uneven intensity and a little discontinuity of edge. The result of our proposed method with 7×7 patch size has the best reconstructed structure (a roof top) as shown in Fig. 5(h). However, there is noticeable blur at texture region (on the top of the roof). This effect may be the result of average updating equation (5) and fixed patch size. In other word, 7×7 patch size may be suitable for reconstructing structure of Fig. 5(b) but the smaller patch size may be needed

for reconstructing texture region. This assumption will be investigating on our future works.

6 Conclusion

This paper proposes a novel iterative version of exemplar-based image inpainting. A main contribution of our proposed method is that the target patch is selected by consideration of structural information. Unlike traditional raster scan approach, whose results are inconsistent, the structures of the damage image tend to be reconstructed perfectly and consistently by our approach. This means that the reconstruction results of the same damaged image with different rotation are unique by our method. Moreover, A fast computation speed of our inpainting method is achieved by an integration of a redesigned random search strategy. Our proposed method outperforms several well-known exemplar-based inpainting on both greedy and iterative approaches, especially on structural region. However, in some cases, the texture nearby the reconstructed structure is flattened. As our assumption, the patch which suitable for reconstruct structure may not be suitable for reconstruct texture. So, designing for inpainting algorithm with adaptive patch size is our future work.

References

1. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 28(3) (August 2009)
2. Bertalmío, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: *SIGGRAPH*, pp. 417–424 (2000)
3. Bertalmío, M., Vese, L.A., Sapiro, G., Osher, S.: Simultaneous structure and texture image inpainting. *IEEE Transactions on Image Processing* 12(8), 882–889 (2003)
4. Chan, T.F., Shen, J.: Local inpainting models and tv inpainting. *SIAM Journal on Applied Mathematics* 62(3), 1019–1043 (2001)
5. Chan, T.F., Shen, J.: Non-texture inpainting by curvature-driven diffusions (cdd). *J. Visual Comm. Image Rep.* 12, 436–449 (2001)
6. Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing* 13(9), 1200–1212 (2004)
7. Efros, A., Leung, T.: Texture synthesis by non-parametric sampling. In: *International Conference on Computer Vision*, pp. 1033–1038 (1999)
8. Kanizsa, G.: *Organization in Vision*. Holt, Rinehart Winston (1979)
9. Komodakis, N., Tziritas, G.: Image completion using efficient belief propagation via priority scheduling and dynamic pruning. *IEEE Transactions on Image Processing* 16(11), 2649–2661 (2007)
10. Kwok, T.H., Sheung, H., Wang, C.C.: Fast query for exemplar-based image completion. *IEEE Transactions on Image Processing* 19, 3106–3115 (2010)
11. Lee, S.Y., Heu, J.H., Kim, C.S., Lee, S.U.: Object removal and inpainting in multi-view video sequences. *International Journal of Innovative Computing, Information and Control* 6(3(B)) (March 2010)

12. Simakov, D., Caspi, Y., Shechtman, E., Irani, M.: Summarizing visual data using bidirectional similarity. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8 (June 2008)
13. Tae-o-sot, S., Nishihara, A.: Exemplar-based image inpainting with patch shifting scheme. In: 17th International Conference on Digital Signal Processing (2011)
14. Wexler, Y., Shechtman, E., Irani, M.: Space-time completion of video. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 463–476 (2007)

Feature Extraction Based on Co-occurrence of Adjacent Local Binary Patterns

Ryusuke Nosaka, Yasuhiro Ohkawa, and Kazuhiro Fukui

Graduate School of Systems and Information Engineering,
University of Tsukuba, Japan

{nosaka, ohkawa}@cvlab.cs.tsukuba.ac.jp, kfukui@cs.tsukuba.ac.jp

Abstract. In this paper, we propose a new image feature based on spatial co-occurrence among micropatterns, where each micropattern is represented by a Local Binary Pattern (LBP). In conventional LBP-based features such as LBP histograms, all the LBPs of micropatterns in the image are packed into a single histogram. Doing so discards important information concerning spatial relations among the LBPs, even though they may contain information about the image's global structure. To consider such spatial relations, we measure their co-occurrence among multiple LBPs. The proposed feature is robust against variations in illumination, a feature inherited from the original LBP, and simultaneously retains more detail of image. The significant advantage of the proposed method versus conventional LBP-based features is demonstrated through experimental results of face and texture recognition using public databases.

Keywords: Image feature extraction, local binary pattern (LBP), co-occurrence, face recognition, texture recognition.

1 Introduction

In this paper, we propose a new Local Binary Pattern (LBP)-based feature by introducing the spatial co-occurrence of adjacent LBPs. LBP-based features, such as LBP histograms, have recently attracted attention as a fundamental technique in the applications of texture recognition, face recognition, and facial expression recognition [1,2,3,4,5], owing to their high robustness to changes in illumination and their efficient computation. The basic idea of the LBP histogram, the focus of this paper, is the representation of entire images as a composition of numerous LBPs, where each LBP is extracted from a local region. LBP was originally designed as a texture description for a local region, called a micropattern [6]. LBP is a binary pattern that represents the magnitude relation between the center pixel of a local region and its neighboring pixels. LBP is obtained by thresholding the image intensity of the surrounding pixels with that of the center pixel. In the LBP histogram, the obtained binary patterns are converted to a decimal number as a label, and a histogram is generated from the labels of all the local regions of the whole image.

Since LBP considers only the magnitude relation between the center and neighboring pixel intensities, LBP is invariant to uniform changes of image intensity over the entire image, making it robust against changes in illumination. This characteristic of LBP has led it to become a standard feature for face recognition and facial expression analysis [7].

Unfortunately, however, spatial relations among the LBPs are mostly discarded during the LBP histogram generation process, because the LBPs are forcedly packed into a single histogram, resulting in the loss of global image information. This suggests that there is still a room for further improvement to the performance of LBP-based features, while retaining invariance to changes in illumination.

To consider the spatial relation among LBPs, we introduce the concept of co-occurrence. Co-occurrence is often used to extract information related to global structures in various local region-based features, e.g. Co-HOG[8], GLAC[9] and Joint Haar-like Features[10]. Although co-occurrence of LBPs can be obtained as a heuristic problem, we introduce a more sophisticated way to obtain the co-occurrences of all combinations of LBPs by using auto-correlation matrices calculated from two considered LBPs. The calculation process will show that the proposed feature is a natural extension of the original LBP in that the proposed feature consists of both the original LBP and the co-occurrence of LBPs.

Fig.1 shows the difference between an LBP histogram and the histogram of the spatial co-occurrence between LBPs. The three image examples are composed of three LBP A and three LBP B patterns, as shown in Fig.1(a). Since the number of LBP A and LBP B patterns in each image are the same, the LBP histograms are generated from the three images coincide with each other, as shown in Fig.1(b). In contrast, the histograms of the spatial co-occurrence extracted from each image are quite different, as shown in Fig.1(c). From this

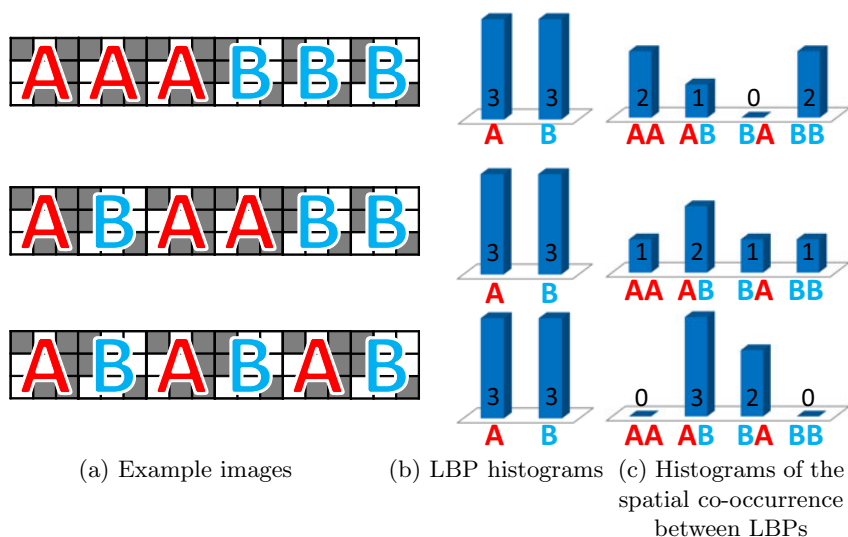


Fig. 1. Difference of LBP histogram and histogram of spatial co-occurrence between LBPs

simple example, we can see that the expression ability of the original LBP is insufficient, and the spatial co-occurrence of LBPs is a valid requirement for realizing a higher expression ability.

The remainder of this paper is organized as follows. In Section 2, we briefly review the LBP algorithm and extend it by introducing the concept of co-occurrence of adjacent LBPs. Section 3 describes evaluation experiments confirming the validity of the proposed feature. Section 4 presents our conclusions.

2 The Proposed LBP-Based Image Feature

2.1 Local Binary Pattern Histograms

LBP indicates a magnitude relation between a center pixel and its neighboring pixels in a micropattern. Fig.2 shows an example of a micropattern and an LBP corresponding to it. The LBP ‘10000111’ is obtained by thresholding 3×3 neighbor pixels with the value of the center pixel, 5. The binary pattern is then converted to its decimal equivalent, 135. LBP histograms are generated from the decimal values of all the micropatterns.

Let I be an image intensity and $\mathbf{r} = (x, y)^T$ be a position vector in I . The LBP $\mathbf{b}(\mathbf{r}) (\in \mathbb{R}^{N_n})$ is defined as follows:

$$b_i(\mathbf{r}) = \begin{cases} 1, & I(\mathbf{r}) < I(\mathbf{r} + \Delta \mathbf{s}_i) \\ 0, & \text{otherwise} \end{cases}, (i = 1, \dots, N_n), \quad (1)$$

where N_n is the number of neighbor pixels and $\Delta \mathbf{s}_i$ are displacement vectors from the position of center pixel \mathbf{r} to neighbor pixels. In the original LBP, these parameters are set as follows: $N_n = 8$, $\Delta \mathbf{s}_i \in \{(\pm \Delta s, \pm \Delta s)^T, (\pm \Delta s, \mp \Delta s)^T, (\pm \Delta s, 0)^T, (0, \pm \Delta s)^T\}$, $\Delta s = 1$. Next, the LBP $\mathbf{b}(\mathbf{r})$ is converted to a decimal number. Finally, the histogram of the LBPs is generated by considering the decimals as labels.

The magnitude relation of intensities is invariant for change such as scalar times intensity in an entire image. In other cases of illumination conditions, the magnitude relation is also robust. Therefore, LBP is robust against illumination variance.

2.2 Co-occurrence of Adjacent Local Binary Patterns

The co-occurrence of adjacent LBPs is defined as an index of how often their combination occurs in the whole image. Here, we explain how to calculate the co-occurrence of LBPs.

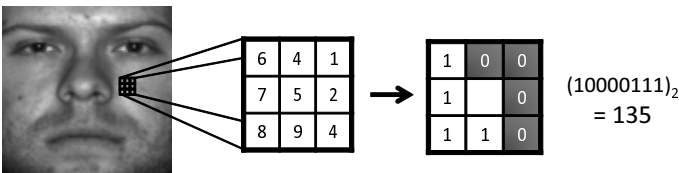


Fig. 2. Original LBP

Fig.3 shows a comparison of the original LBP and the co-occurrence of LBPs. The original LBP shown in Fig.3(a) can represent only a simple image pattern. On the other hand, the combination of multiple LBPs can represent various image patterns derived from more complicated surfaces, as shown in Fig.3(b).

In the proposed feature, the number of possible combinations of LBPs is significantly greater than that of the original LBPs. It is therefore difficult to use a rule-based program to compute the co-occurrence of all combinations when there are many types of LBPs. Instead of using a heuristic program, we introduce an auto-correlation matrix as an effective method of calculating the co-occurrence of LBPs. First, although the original LBP uses eight neighbor pixels of a given center pixel, we modify the LBP configuration to consider two sparser configurations, thereby reducing computational cost. One configuration is LBP(+), which considers only two horizontal and two vertical pixels, as shown in Fig.4(a). The other configuration is the LBP(\times), which considers the four diagonal pixels shown in Fig.4(b). In the LBP(+), the parameters are set as follow: $N_n = 4$, $\mathbf{s}_i \in \{(\pm\Delta s, 0)^T, (0, \pm\Delta s)^T\}$. In the LBP(\times), the parameters are set as follow: $N_n = 4$, $\mathbf{s}_i \in \{(\pm\Delta s, \pm\Delta s)^T, (\pm\Delta s, \mp\Delta s)^T\}$.

Next, in order to effectively calculate the co-occurrence of LBPs, each LBP is converted to vector $\mathbf{f}(\in \mathbb{R}^{N_p})$, which is defined as follows:

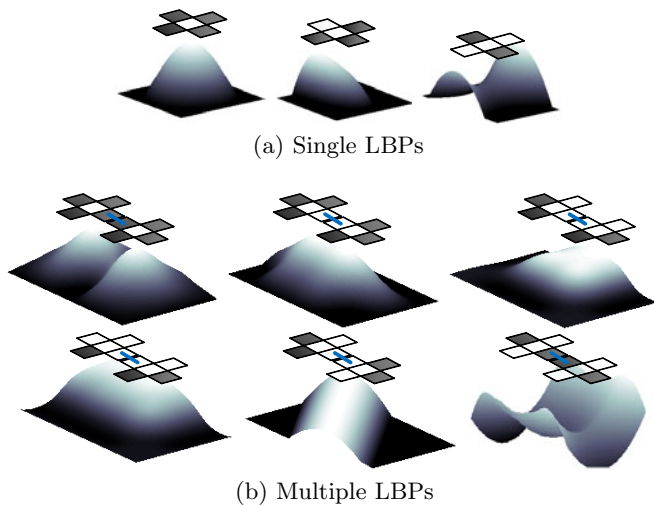


Fig. 3. Comparison of single and multiple LBPs

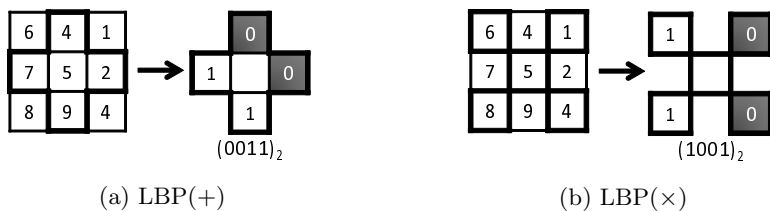


Fig. 4. Sparse LBPs

$$f_i(\mathbf{r}) = \delta_{i,l(\mathbf{b}(\mathbf{r}))}, \tag{2}$$

where, N_p is the number of all the possible LBPs, $N_p = 2^{N_n}$ for example, in the original LBP, $N_p = 2^8 = 256$, N_n is the number of neighbor pixels, $\delta_{i,j}$ is Kronecher’s delta, and $l(\mathbf{b}(\mathbf{r}))$ is the label of $\mathbf{b}(\mathbf{r})$.

To effectively calculate the co-occurrence of LBPs, we consider the $N_p \times N_p$ auto-correlation matrix defined by the following equation:

$$\mathbf{H}(\mathbf{a}) = \sum_{\mathbf{r} \in I} \mathbf{f}(\mathbf{r}) \mathbf{f}(\mathbf{r} + \mathbf{a})^\top, \tag{3}$$

where \mathbf{a} is the displacement vector from the reference LBP to its neighbor LBP. The element $H_{i,j}(\mathbf{a})$ of Eq.(3) indicates the number of pairs of adjacent LBP i LBP j . After shift-equivalent patterns are removed, \mathbf{a} is set as follows: $\{(\Delta r, 0)^\top, (\Delta r, \Delta r)^\top, (0, \Delta r)^\top, (-\Delta r, \Delta r)^\top\}$. Fig.5 shows all the configuration patterns of \mathbf{r} and $\mathbf{r} + \mathbf{a}$ in Eq.(3).

Next, we explain characteristics of the proposed feature. Although the proposed feature has high dimensionality ($4N_p^2$), the computational cost is low due to the sparsity of the LBP. The original LBP histogram can be obtained as the summation of the column vectors of the matrix $\mathbf{H}(\mathbf{a})$ defined by Eq.(3). This means that the proposed feature retains the original LBPs along with co-occurrence information, making it a natural extension of the original LBP.

Fig.6 shows the flow of extracting the proposed feature from an image. The example image has four LBPs (Fig.6(a)). The labels of these LBPs are 2, 8, 9, and 14, respectively. In the case of the displacement vector $\mathbf{a} = (\Delta r, 0)^\top$, there are two LBP pairs ($\{\text{upper left, upper right}\}$ and $\{\text{lower left, lower right}\}$) in the image. Since the labels are (2, 14) and (8, 9), the elements corresponding to these labels in Eq.(3) are set to 1 and other elements are set to 0. For other displacement vectors: $\mathbf{a} = (0, \Delta r)^\top, (\Delta r, \Delta r)^\top$ and $(-\Delta r, \Delta r)^\top$, an auto-correlation matrix $\mathbf{H}(\mathbf{a})$ is similarly generated as shown in Fig.6(b).

Fig.7 shows the process flow of proposed feature. Firstly, LBPs are extracted from the input image as shown in Fig.7(a). Next, we compute four $N_p \times N_p$ auto-correlation matrices of spacial co-occurrences of adjacent LBPs, $\mathbf{H}(\mathbf{a})$, as shown in Fig.7(b). Finally, these matrices are vectorized and combined to a $4N_p^2$ -dimensional feature vector \mathbf{z} (Fig.7(c)).

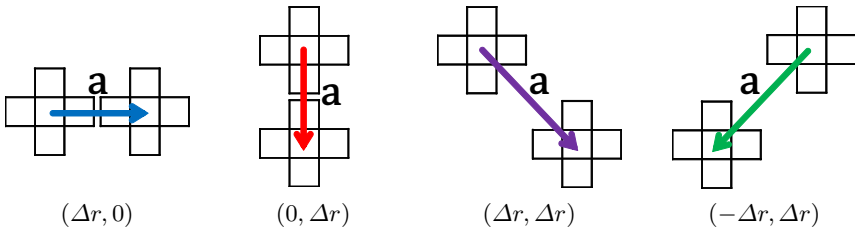


Fig. 5. Configuration patterns of proposed feature

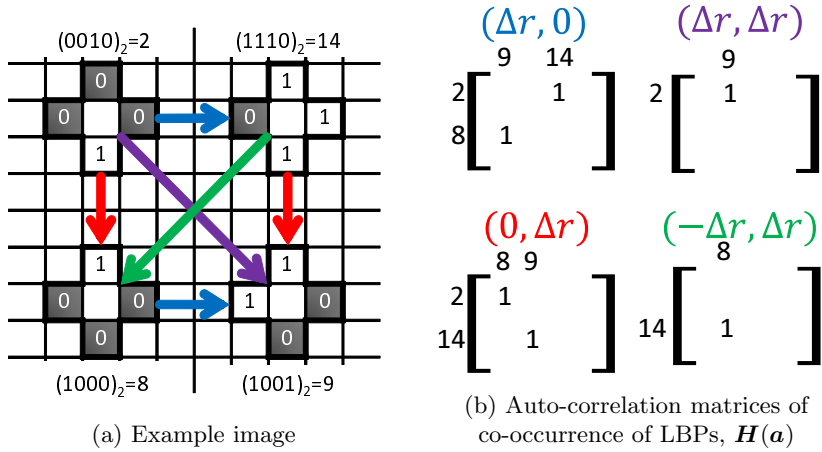


Fig. 6. Example of obtaining proposed feature

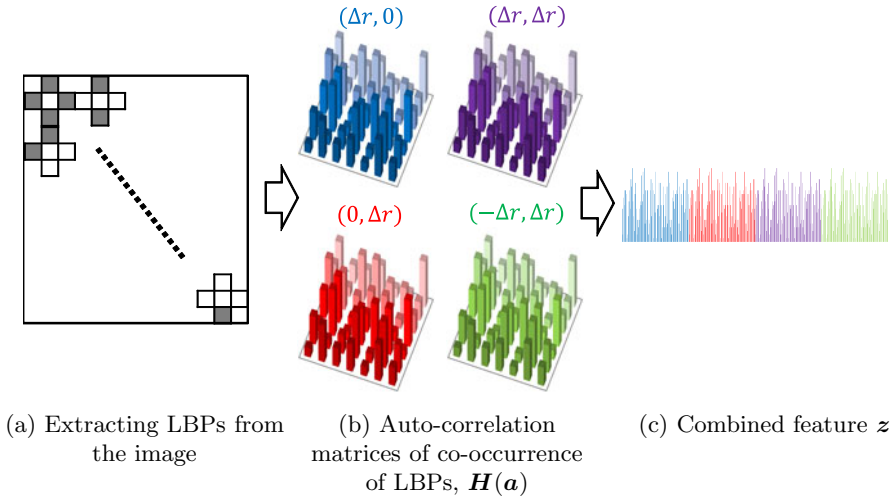


Fig. 7. Process flow of proposed feature

3 Experiments and Considerations

To confirm the proposed method’s validity, we applied it to face recognition and texture recognition tasks.

3.1 Face Recognition Experiment

Setup. In this experiment, we evaluated the proposed feature using the Extended Yale Face Database B [11]. Fig.8 shows examples of images included in the dataset. The dataset contains the faces of 38 subjects under 64 variations in



Fig. 8. Examples from the Extended Yale Face Database B

illumination for each. All images are frontal face images cropped to 168×168 pixels. We resized the images to 88×88 pixels for this experiment. Images with frontal lighting were used as a training set (one image per person). The remaining images were used as a testing set (63 images per person).

We compared the proposed feature with a raw image feature, a Gabor image feature [12], and the LBP histogram. The raw image feature was obtained by vectorizing the input image.

For the LBP histogram we prepared three types of features, differing in the selection of surrounding pixels: 3×3 neighbor pixels (original LBP), LBP(+) and LBP(\times). For the proposed feature, we prepared two types of features, differing in the selection of surrounding pixels: LBP(+) and LBP(\times).

The parameters of the proposed feature were changed as follows: $\Delta s = 1, \dots, 5$ and $\Delta r = 1, \dots, 20$. The best correct rate among the results was regarded as the final result.

The image was divided into multiple subregions. Four types of divisions (1×1 , 2×2 , 4×4 , and 8×8) were performed. The features extracted from these subregions were integrated into a final feature \mathbf{z} . Therefore, for each division, the dimension of the final proposed feature \mathbf{z} is $4N_p^2 \times 1$, $4N_p^2 \times 2^2$, $4N_p^2 \times 4^2$, $4N_p^2 \times 8^2$, respectively. The region division described above was not performed for the raw image feature and the Gabor image feature.

The nearest neighbor method with L_1 norm was used as a classifier. The L_1 norm is usually used as the similarity between two histograms [13], since it has a similar characteristic to the histogram intersection defined by $S(\mathbf{H}_1, \mathbf{H}_2) = \sum_i \min(H_{1i}, H_{2i})$.

Results. Fig.9 shows the results of the experiment. From these results, we can confirm the LBP-based feature's robustness against variations in illumination. In contrast, performances of the raw image feature and the Gabor image feature were poor, due to their sensitivity to illumination. Their performances could not be improved even when considering the co-occurrence.

The results show that the histogram feature of LBPs outperform other features. In particular, the performance of the proposed feature with co-occurrence of LBPs is remarkable. It has achieved the best performance using parameters $\Delta s = 1$, $\Delta r = 3$, and 8×8 division. We can see that increasing number of division increases the performance due to keeping spacial information.

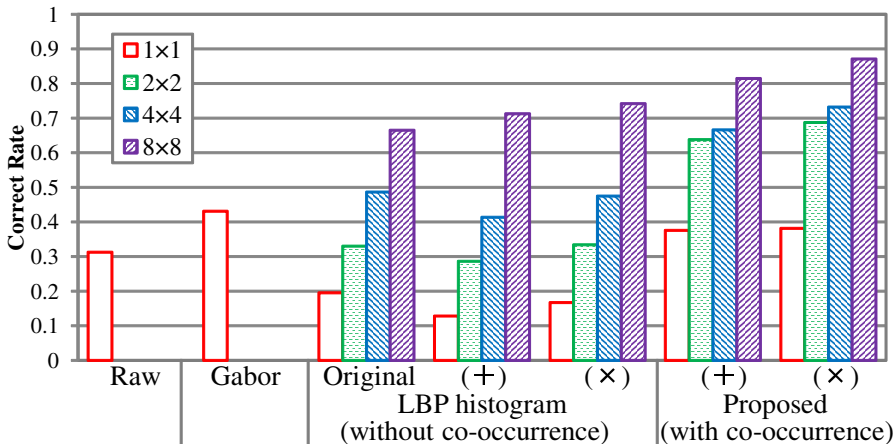


Fig. 9. Results of face recognition experiment

Moreover, we can also see that the performance of the proposed feature depends on the configuration of LBP. When the division number was set to a large value, the performances of both the LBP(+) and LBP(×) were superior to the original LBP. The reason for this result can be explained as follows. The dimension of the original LBP feature is higher than that of the LBP(+) and LBP(×). Therefore, the distribution of the original LBPs tends to be too sparse and unstable as the division number increases.

3.2 Texture Recognition Experiment

Setup. In this experiment, we evaluated the proposed method using Outex_TC from the public database Outex [14]. Table 1 shows the details of the dataset. Fig.10 shows examples of images from the dataset.

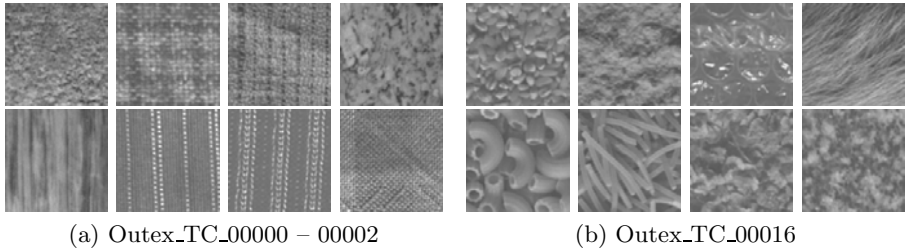
Outex_TC_00000 – 00002 contains grayscale images of 24 textures at different sizes. Outex_TC_00016 contains 319 textures. The average image intensity value is normalized to 128, with a standard deviation of 20.

The images were randomly split between training and testing sets. This division was repeated 100 times to produce 100 evaluation sets. The average of all correct rates over 100 iterations was defined as the final rate.

The proposed feature employed $\Delta s = 1, \dots, 3$ and $\Delta r = 1, \dots, 5$, and the best correct rates were used as the reported results. The dimension of the final

Table 1. Outex database details

Outex ID	Classes	Image sizes	Training/Testing images
Outex_TC_00000	24	128 × 128	10
Outex_TC_00001	24	64 × 64	44
Outex_TC_00002	24	32 × 32	184
Outex_TC_00016	319	128 × 128	10

**Fig. 10.** Outex database examples**Table 2.** Results of the texture recognition experiment

Outex ID	LBP histogram			Proposed	
	Original	(+)	(×)	(+)	(×)
Outex_TC_00000	0.996	0.986	0.989	0.999	0.999
Outex_TC_00001	0.985	0.930	0.948	0.989	0.989
Outex_TC_00002	0.871	0.721	0.742	0.906	0.915
Outex_TC_00016	0.783	0.686	0.708	0.830	0.820

proposed feature \mathbf{z} is $4N_p^2$. The L_1 nearest neighbor method was used as a classifier.

Results. Table 2 shows the results of the experiment. The results confirm a significant advantage of the proposed feature against the LBP histogram features, which are not considering the co-occurrence. The proposed feature with parameters $\Delta s = 1$, $\Delta r = 2$ achieved the best performance among all the features.

In contrast to the previous experiment, the original LBP was better than the LBP(+) and LBP(×) in the case that co-occurrence was not considered. This is because the subregion size used in this experiment was better suited to the original LBP, as compared with other LBPs.

4 Conclusion

We have proposed a novel image feature based on the spatial co-occurrence of micropatterns, which are represented by Local Binary Pattern (LBP). The conventional LBP-based features as represented by the LBP histogram still has room for performance improvements. In particular, expression ability for a given image can be improved, since all LBPs are simply summed into a single histogram, thereby discarding spatial relations among the LBPs and the rich image information they contain. To improve their performance, we introduced the extension of original LBP by considering the co-occurrence of adjacent LBPs, measuring co-occurrence with an auto-correlation matrix generated from multiple LBPs. The proposed feature is robust against variations in illumination, because it only depends on the magnitude relation between a center pixel and its surrounding pixels. Experimental results of face and texture recognition tasks using public

databases have demonstrated a significant advantage of the proposed feature against conventional LBP-based features.

References

1. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 971–987 (2002)
2. Zhang, W., Shan, S., Gao, W., Chen, X., Zhang, H.: Local Gabor Binary Pattern Histogram Sequence (LGBPHS): A Novel Non-Statistical Model for Face Representation and Recognition. In: *Proc. IEEE International Conference on Computer Vision*, vol. 1, pp. 786–791 (2005)
3. Zhao, G., Pietikäinen, M.: Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 915–928 (2007)
4. Lei, Z., Liao, S., He, R., Pietikainen, M., Li, S.: Gabor volume based local binary pattern for face representation and recognition. In: *Proc. IEEE Conference on Automatic Face and Gesture Recognition*, pp. 1–6 (2008)
5. Zhang, B., Gao, Y., Zhao, S., Liu, J.: Local derivative pattern versus local binary pattern: face recognition with high-order local pattern descriptor. *IEEE Transactions on Image Processing* 19, 533–544 (2010)
6. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition* 29, 51–59 (1996)
7. Ahonen, T., Hadid, A., Pietikäinen, M.: Face description with local binary patterns: Application to face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 2037–2041 (2006)
8. Watanabe, T., Ito, S., Yokoi, K.: Co-occurrence Histograms of Oriented Gradients for Pedestrian Detection. In: *Proc. the 3rd IEEE Pacific-Rim Symposium on Image and Video Technology*, pp. 37–47 (2009)
9. Kobayashi, T., Otsu, N.: Image Feature Extraction Using Gradient Local Auto-Correlations. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part I*. LNCS, vol. 5302, pp. 346–358. Springer, Heidelberg (2008)
10. Mita, T., Kaneko, T., Stenger, B., Hori, O.: Discriminative feature co-occurrence selection for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 1257–1269 (2008)
11. Lee, K., Ho, J.: Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 684–698 (2005)
12. Liu, C., Wechsler, H.: Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. *IEEE Transactions on Image Processing* 11, 467–476 (2002)
13. Swain Jr., M., Ballard, D.: Color indexing. *International Journal of Computer Vision* 7, 11–32 (1991)
14. Ojala, T., Mäenpää, T., Pietikäinen, M., Viertola, J., Kyllönen, J., Huovinen, S.: Outex - new framework for empirical evaluation of texture analysis algorithms. In: *Proc. IEEE International Conference on Pattern Recognition*, pp. 701–706 (2002)

Natural Image Composition with Inhomogeneous Boundaries

Dong Wang¹, Weijia Jia¹, Guiqing Li², and Yunhui Xiong²

¹ City University of Hong Kong, China

² South China University of Technology, China
donwang@student.cityu.edu.hk,
wei.jia @cityu.edu.hk,
{ligq,yhxiong}@scut.edu.cn

Abstract. Image composition usually floods the composition region of a target image with the same shape as a source image patch. To achieve seamless transition effect, the tone of the boundary in the target image is then transferred to the interior region of the source patch. Traditional approaches usually fail to work for the case that the corresponding boundaries of target and source images don't match well because the tone transformation of all pixels on the boundary are equally propagated to the inner region. This paper presents a new image composition technique based on discrete mean value coordinates(DMVC), which supports the transition of tone transformation of part selected not all pixels on the boundary to the inner region. The approach works as follows. It firstly selects boundary pixels having good matching. The new color of inner pixels is then calculated using DMVC according to those selected pixel pairs from the source and target boundaries. Matting technique is finally introduced to compose the new pixels to the target image. Experiments show that the proposed approach can obtain reasonable results for examples with inconsistent boundaries between source and target images.

Keywords: Image composition, Discrete mean value coordinates, Image editing, Edit propagation.

1 Introduction

Image composition as an interactive image editing operation plays an important role in image processing. Its purpose is to generate a new image by merging the selected image patches from one or more images into one base image naturally and seamlessly. It is widely used in moving making, photo processing, and web design etc. In general, there are two kinds of image composition techniques: image cloning and image blending [1]. For image cloning each pixel in the new composite image is from either the source image or the target image; while image blending deals with a blending operator about alpha between source image pixel and corresponding target pixel. In order to achieve seamless composite effect, we adopt image blending technique by combing alpha matting method.

A trivial image composition is to copy the foreground of the source image patch to some region of the target image without any further processing. It is hard for this method to produce seamless composition results with a unified color tone due to the fact that the target image and source image generally have different hue style. On the other hand, it is not easy for users to exactly extract the foreground of the source image. It is expected that the foreground objects can be quickly specified by sketching a rough region boundary in which not only the foreground objects but also some background pixels are contained. In fact, it is useful to include some background pixels around the objects to be cloned as they can act as the role of a smooth transition from target image to these objects. However, it will exhibit artifacts in the composed image if the corresponding background textures of source and target images are severely different. Moreover, if the selected region is too large, some important content of the target image is possibly deteriorated [2].

As illuminations of the source region and the target image are usually inconsistent, their color tones are generally unmatched and hue style transfer is therefore necessary. Gradient-based methods are a natural choice for conducting such a task [3]. This type of methods generally force the gradient of the composite region accordant with that of the source region while interpolating the boundary pixels from the target image. It finally results in a Laplace equation which can be discredited as a large sparse linear system. As an approximation of harmonic functions, Fleisman et al. employ mean value coordinates (MVC) to interpolate the boundary mismatch between target and source images [4]. This yields a more efficient cloning scheme. The composite boundary plays an important role for both methods. The effectiveness of image composition depends on the selection of boundaries considerably. If the texture and color of the source and target boundaries are consistent, they can achieve satisfactory results. If inconsistent, however, it may bring unrealistic artifacts. This indicates that it is necessary to carefully consider the boundary problem.

A possible way is to automatically find a closed path enclosing the objects to be cloned in the source image and a corresponding region in the target image such that the mismatch is minimized. However, users may usually be willing to specify the boundary of the source objects and the position they should be pasted in the target image. Chen et al. [5] attacked the problem by classifying boundary pixels into consistent and inconsistent sets and employ common and blending boundary conditions, respectively. Ding and Tong [6] further improved the Poisson image composition method by weighting the gradient error which finally yielded a MVC-like scheme. However, both methods do not support controllable boundary matching.

To attack the problem, we put the boundary pairs between source and target regions into four classes: whole-to-whole, part-to-whole, whole-to-part, and part-to-part when propagating the boundary mismatch to the inner source region, where A-to-B indicates that the color tone of A pixels from the source boundary should be transferred to the tone of B pixels from the target boundary. What kind of pairs is chosen depends on the mismatch degree and we provide an

automatic method based on a matching criteria to select a boundary matching style. Of course, it can also be specified by users.

It should be pointed out that the last three classes result in open boundaries. It is difficult to directly apply gradient-based/MVC-based methods to these cases as they require a closed boundary. This motivates us to adopt a DMVC interpolation to propagate the mismatch. Our contributions are as follows: (1) A new composition algorithm based on DMVC is devised, which can deal with the propagation of open boundaries; (2) It can generate a reasonable composition even when the boundaries of target and source regions do not match well while it is difficult to do this for previous methods.

2 Related Work

There are two basic image composition approaches: alpha blending and gradient-based method [6]. Alpha blending approaches are mainly focused on alpha calculation. In [7], the matte is directly reconstructed from matte gradient field by solving Poisson equation with boundary information from a user-supplied trimap. [8] proposed a closed-form matting approach by explicitly deriving a cost function from local smoothness assumptions on the foreground and background colors, where each foreground and background is a linear mixture of two colors over a small neighborhood around each pixel. [9] introduced a graph cut technique to develop a method of interactive foreground extraction and alpha-matting of color images. The algorithm builds a Gaussian mixture model of color image based on a given trimap, and iteratively alternates between graph cuts and the updated model parameters until it satisfies a specified convergence criterion. If we divide the composition process into two parts: cut and paste, alpha blending mainly concerns about the cut, how to get the expected region.

Gradient-based methods pay more attention to the pasting effect. [3] gave a Poisson image editing method where the gradient field inside the cloned region is taken from the source image and the Dirichlet boundary conditions for the equation is defined by the boundary of the cloned region where the pixel color is from the corresponding target image. For gradient domain techniques need to solve a large sparse linear system, a number of works proposed fast Poisson solvers for various scenarios [10,11] and for solving the Poisson equation on the GPU [12]. [4] gave a different fast algorithm by introducing mean value coordinates(MVC) image cloning. The new color of cloned region is the sum of the base color from the source image patch and the offset. The offset along the boundary of the cloned region is the difference between the corresponding pixel of the source image and the target image, while the offset of each interior pixel is given by a weighted combination of values along the boundary.

To preserve the color fidelity of the foreground objects of source images, [1,6] gave their different solutions. Based on the Poisson image editing framework, [1] proposed a variational model that considers both the gradient constraint and the color fidelity. They did a soft segmentation to the source image to obtain the foreground and the new gradient of cloned region is the weight sum of the

corresponding gradient of the source image and the target image with the soft segmentation as weight. [6] discussed a content-aware image blending technique. Based on MVC image cloning, the offset is controlled by a coefficient related to the alpha, where they can preserve the color of foreground region unchanged.

When the corresponding boundary of source image and target image do not match, it may generate blending artifacts in the composite image by gradient-based methods, as discussed in [2] and [13]. In [2] they proposed a new objective function to compute an optimized boundary condition. A shortest closed-path algorithm is designed to search for the location of the boundary. But their method may produce local optimization. In [13] they improved standard gradient-domain compositing technique. They redefine the boundary gradients to ensure the produced gradient field is nearly integrable and spread the residuals to those regions where they are less conspicuous. They do not overcome but disperse the bleeding artifacts.

We cast image composition as a discrete MVC interpolation problem which diffuses the difference between pixels from source and target boundaries. The method does not pose any special requirements on initial conditions. The approach can produce more natural composition results than gradient-based and MVC-based composition techniques for the case with complicated boundary color tones.

3 Discrete MVC Image Composition

In MVC-based image cloning [4], a new pixel in the composite region is obtained by adding an offset to the corresponding pixel of the source region. The offset on boundary pixels is defined as the difference of the corresponding boundary pixels. The offset inside composite region is calculated by an MVC-based interpolatory function interpolating the boundary offsets. In our discrete MVC image composition, we allow offsets of a portion of boundary pixels to be diffused while other boundary pixels are treated using blending technique. In this section, we firstly describe our problem formulation, then derive the discrete MVC offset propagation, and finally outline the composition algorithm.

3.1 Nomenclature

Suppose a source image $\iota_s : \mathbb{R}^2 \rightarrow \mathbb{R}^d$ and a target image $\iota_t : \mathbb{R}^2 \rightarrow \mathbb{R}^d$ are given, where d is the color feature dimension. A source image region $\Omega_s \subset \mathbb{R}^2$ to be cloned onto $\Omega_t \subset \mathbb{R}^2$ is selected by interactively specifying its closed boundary $\partial\Omega_s$. The cloned region $\Omega_t \subset \mathbb{R}^2$ in the target image ι_t is uniquely determined by a translation $o(o_x, o_y) : \Omega_t = \{q' | q' = q + o, q \in \Omega_s\}$. Let $P_s \subset \partial\Omega_s$ be a set of pixels on the boundary of the source region, and $P_t \subset \partial\Omega_t$ the corresponding pixel set on the boundary of the target image. Furthermore, denote the composition image by $\iota_c : \mathbb{R}^2 \rightarrow \mathbb{R}^d$. Finally, $\alpha : \mathbb{R}^2 \rightarrow \mathbb{R}$ is a mask such that

$$\alpha(p) \begin{cases} = 0, p \in \Omega_b, \\ = 1, p \in \Omega_f, \\ \in (0, 1), \textit{otherwise.} \end{cases}$$

where $\Omega_b(\subset \Omega_s)$ is the background and $\Omega_f(\subset \Omega_s)$ is the foreground, which will be used as blending weight. To achieve good composition quality, α should be the sampling of a smooth function for which α maps calculated using image matting methods [7,8,9] are good candidates.

3.2 Discrete MVC Interpolatory Function

For given planar polygon P and function f defined on the boundary of the polygon, the MVC interpolation function on the plane is defined as [14,15,16]

$$f(p) = \frac{\int_{q \in \partial P} \omega(p, q) f(q) dq}{\int_{q \in \partial P} \omega(p, q) dq},$$

where $\omega(p, q) = \frac{1}{\|p-q\|}$ which we replace with $\omega(p, q) = \frac{1}{\|p-q\|^2}$ for efficiency consideration.

In our setting, only a set of discrete pixels are known for interpolation, therefore we introduce a so-called discrete MVC interpolatory function

$$f_P(p) = \frac{\sum_{q \in \partial P} \omega(p, q) f_P(q)}{\sum_{q \in \partial P} \omega(p, q)},$$

where P is a discrete point (pixel) set.

3.3 Image Composition

Now we can produce an intermediate image ι_i by diffusing the color offsets between P_s and P_t in order to transfer the color tone of Ω_t to ι_s using the discrete MVC interpolatory function

$$\iota_i(p) = \iota_s(p) + f_P(p), \forall p \in \mathbb{R}^2$$

with $f_P(q) = \iota_t(q + o) - \iota_s(q)$ for $q \in P_s$ which is the color difference between target and source images.

After this processing, ι_i will have a similar appearance to that determined by pixels in P_t . We then embed the corresponding region in ι_i to Ω_t of the target image smoothly with alpha blending:

$$\iota_c(q) = \alpha(q - o)\iota_i(q - o) + (1 - \alpha(q - o))\iota_t(q), \forall q \in \mathbb{R}^2$$

4 Boundary Matching Analysis

In image composition, the boundary, as initial conditions, its color in the source image patch is replaced by that of the corresponding target region. And the color inside the composition region should be renewed following the change on boundaries. But in some cases, the source boundary and the target boundary may be greatly different. Just like the case in Fig. As an example, let us see 1(a) in which the source boundary consists of two parts of pixels: one is a leaf part and the other is a sky part. If we only select leaf pixels of the source region to compute edit propagation, the flower should follow a transformation between leaves instead of sky to leaves, and therefore receive little change. But if the sky part of pixels is chosen, the flower should greatly be changed following the difference between between sky pixels and the leaf pixels. So in order to make the composite image more natural and realistic, we need to select the proper initial conditions from the boundary by analyzing the source image and the target image (see 1(c) for example). We categorize the matching types of the source and target boundaries into four classes: (1) whole-to-whole, (2) part-to-whole, (3) whole-to-part, and (4) part-to-part, where the former refers to the boundary status of the source region while the latter stands for the boundary status of the target boundary. For the first class composition that the two boundaries completely match, previous methods can produce ideal results. All other three cases, which belong to partially matching, are be discussed in the following subsections.

4.1 Part-to-Whole Matching Composition

Let's observe Fig. 1 again. We want to replace the red flower in the target image (1 (b)) with the yellow rose in the source image. To avoid the inconsistent matching between whole-to-whole boundaries, we only select leaf pixels from the source image which can match the boundary pixels of the target image very well and generate the composition image by propagating the difference between the selected source pixels and the corresponding target image pixels to the inner pixels of the source patch. To extract those pixels quickly, some reference pixels are provided interactively. We sketch the contour from the leaves and choose a portion of pixels (one percent of the contour pixels for example) as reference pixels. The remaining pixels of the contour are classified by a matching method. The matching idea is that similar color appearance belongs to the same class, and the matching function is defined as

$$s_i = \max_j (\exp(-\|f_i - f_j\|^2) / \delta_f), \quad (1)$$

where i and j are both boundary pixels in the source image. i is the i th pixel whose matching value s_i will be calculated and j is the reference pixel. The meaning of the function is that pixel i matches with all reference pixels. If the maximum value $s_i > \varepsilon$ ($\varepsilon = 0.9$ in our experiments), the corresponding pixel will be selected as the initial condition to calculate the new composite image. The selected pixels are shown in Fig. 1(c) for the example in Fig. 1.

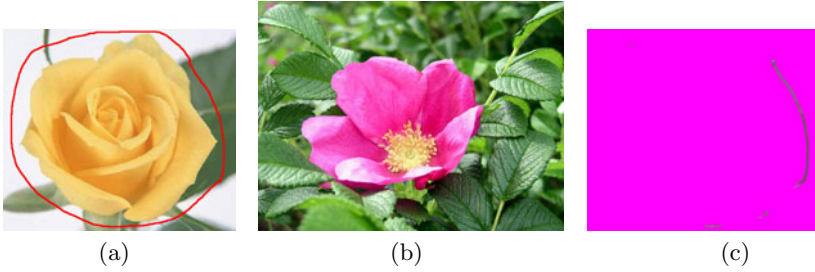


Fig. 1. A part-to-whole example: (a) original image, and selected region, (b) target image, (c) selected boundary pixels as initial conditions for image composition

4.2 Whole-to-Part Matching Composition

In Fig. 2, we hope to move the sand pyramid in the source image into the right top corner of the target image, as shown in Fig. 2(a). The sand pyramid is only surrounded by a kind of pixels, sand pixels, in the source image patch, but there are two kinds of pixels, sand pixels and sky pixels in the target boundary. Noticing that sand pixels and sky pixels are very different in color, the sky pixels will play a negative role in composition if we use the whole boundary pixels to calculate the result as shown in Fig. 6 (d). In this case, we need to extract the pixels from the target boundary which corresponds to sand. Automatic discrimination is difficult too. We again address the problem by manually selecting some reference pixels, (see the red curve in Fig. 2(b) as an example). The difference is that the reference pixels should be selected from the target boundary but not the source boundary. Figure 2(c) illustrates the matching result.

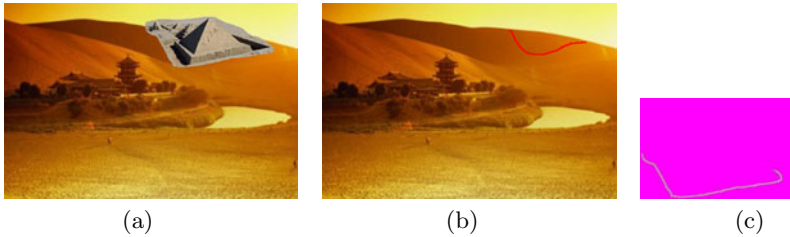


Fig. 2. A whole-to-part example: (a) original image patch and its composition position, (b) reference pixels for boundary pixel selection (marked with red), (c) selected boundary pixels

4.3 Part-to-Part Matching Composition

Figure 3 shows a part-to-part boundary matching example. We want to put two people in on the beach under the coco of the target image. It is a sunning day and there is shadow. The shadow map is created (at the right top of Fig. 3(a))

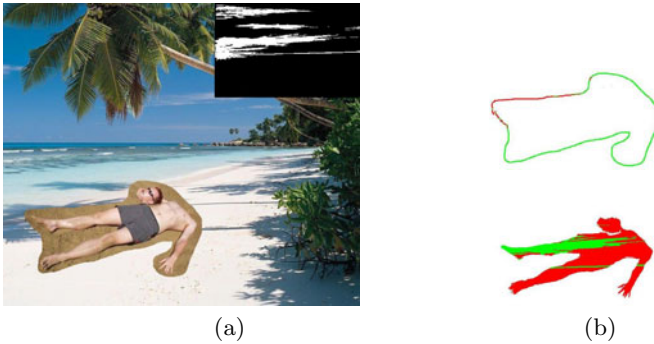


Fig. 3. A part-to-part example:(a) the shadow, original image patch and its composition position, (b) up: the divided boundary, shadow boundary red and the remaining green; bottom: the divided foreground, shadow region red and the remaining green

using Photoshop, which is a binary image. The shadow divides the people and the boundary in two cases, as Fig. 3(b). One is in shadow (marked with red) and the other is under the sun (marked with green). From Fig. 3(b), we can notice each case is made up of several parts, which presents complex boundary. The two cases will be dealt with separately as two composition regions.

5 Experiments and Discussion

In this section, we show our composite results by the proposed approach with a wide variety of source and target images and synchronously compare it with direct pasting, Poisson image editing and error-tolerant image compositing approaches. In our experiments, besides the initial conditions discussed in Section 4, we also need an alpha map for the source image with copied object as foreground in addition to the source and target images. With all these as input, the composition can be computed automatically and quickly.

Figure 4 depicts an example of part-to-whole type corresponding to images in Fig. 1. An alpha map is generated by Levins method [8] for the source image. The illumination of the target image is stronger than that of the source image. Fig. 4(c) presents the result by directly pasting the rose of the source image onto a region of the target image. It is obvious that color tone of the rose cannot match that of the background well as its illumination is weak. Figure 4(d) gives the result by Poisson editing. The color tone of the rose changes unevenly in this case and result exhibits obvious inharmony between the left top corner and the right bottom corner of the composite region. This is caused by unequal color difference between the source image and the target image boundaries. This phenomenon is improved in Figure 4(e) by error-tolerant image compositing method, which disperses the difference into the rose. However, the color appearance of the main object (rose) is almost completely different from its natural color in both results. Figure 4(f) illustrates our result in which only illumination is changed and the

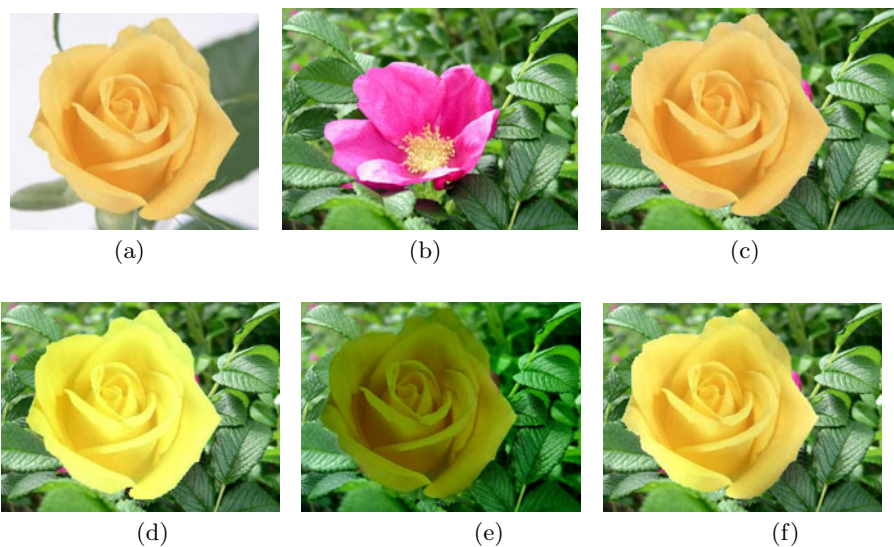


Fig. 4. A part-to-whole image composition example:(a) original image, (b) target image, (c) direct cut-and-paste,(d) Poisson method, (e)Tao's method [13], (f)our method

natural color of the rose is preserved very well. Figure 5 presents another part-to-whole example.

Figure 6 deals with the whole-to-part case. The color appearance of the sand of the target image is obviously different from that of the source image region whose boundary is homogeneous. Fig. 6(c) presents the result by the direct pasting technique. Figure 6(d) gives the result synthesized by the Poisson image editing technique. Since a closed boundary is demanded, the color tone of the top of sand pyramid is closer to the sky color tone and far away from the sand color tone. It is unrealistic. To alleviate the artifact, the boundary gradients are adjusted by error-tolerant image compositing method. The result shown in Fig. 6(e) is created by this improved method. It disperses the color difference



Fig. 5. Another part-to-whole image composition example:(a) original image, (b) direct cut-and-paste,(c) Poisson method, (d) our method

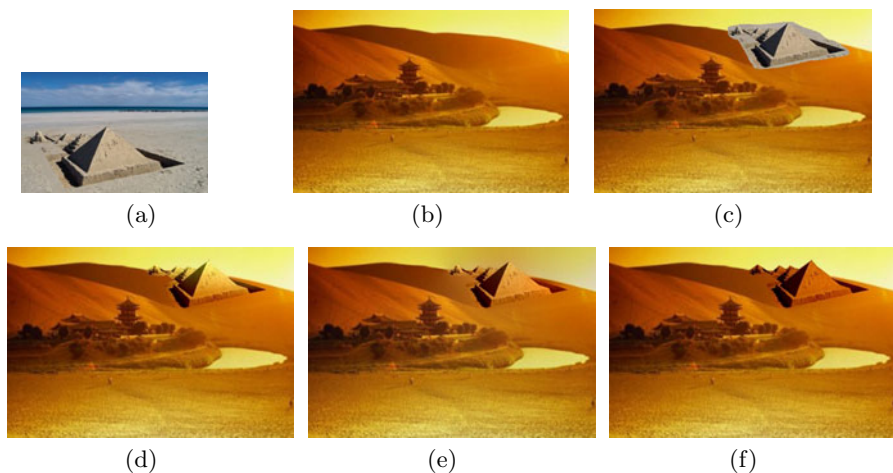


Fig. 6. A whole-to-part image composition example:(a) original image, (b) target image, (c) direct cut-and-paste,(d) Poisson method, (e)Tao's method [13], (f)our method

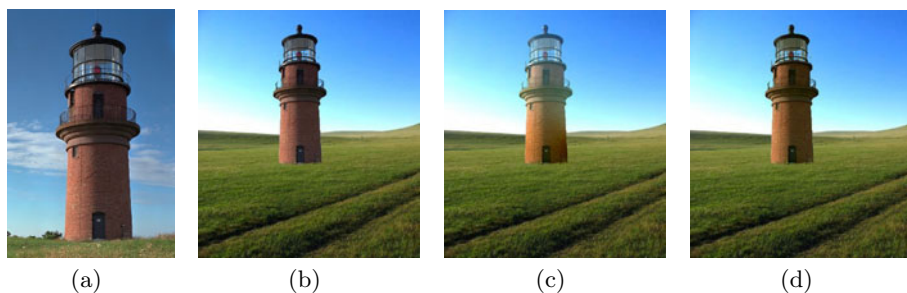


Fig. 7. Another whole-to-part image composition example:(a) original image, (b) direct cut-and-paste,(c) Poisson method, (d) our method

between the sky and the sand into the whole sand pyramid and therefore changes the sky color tone. Our approach clearly outperforms both methods as shown in Fig. 6(f). The color tone of the sand pyramid is well consistent with that of the target region with color tone of the left area deeper than that of the right area. Furthermore, its color tone has nothing with that of the sky color. Figure 7 present another example of this case.

Finally, Figure 8 demonstrate is a part-to-part example corresponding to images in Fig. 3. The color tone between the source image and the target image is very different for the latter includes a set of shadow pixels while the former hasn't. Again, the Poisson image editing composition cannot clearly reproduce the shadow on the man body as shown in Fig. 8 (b) due to the whole boundary color difference propagation while our result can nicely reproduce the sunshine

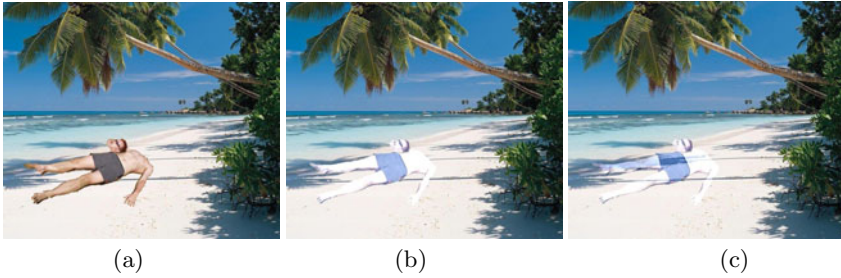


Fig. 8. Image composition with shadow:(a) direct cut-and-paste,(b) Poisson method, (c) our method

and shadow as shown in Fig. 8(d) attributing to intelligently selecting boundary pixels for propagation.

6 Conclusion

We present a discrete mean value coordinates based image composition technique which can generate natural composition images even if the boundaries of the source and target images cannot match well. Compared with previous methods, our method wisely selects an open boundary as the initial condition, and therefore can provide a good solution to the composition of some complex scenes with inhomogeneous boundary color tones.

In our current implementation, our method needs interaction in selection of boundary pixels. An automatic technique is expected in the future work by analyzing boundaries of the given source image patch and the corresponding target region, for example. In fact, there exist many special complex cases to be handled in image composition, such as shadow treatment, texture structure preserving or transferring, and so on. Finally, it is also interesting to investigate methods evaluating the pros and cons of existing composition techniques.

Acknowledgement. The work is partially supported by NSFC (60973084), NSF of Guangdong (915106410-1000106), and Fundamental Research Funds for the Central Universities (2009zz0016).

References

1. Yang, W., Zheng, J., Cai, J., Rahardja, S., Chen, C.: Natural and seamless image composition with color control. *IEEE Transactions on Image Processing* 18(11), 2584–2592 (2009)
2. Jia, J., Sun, J., Tang, C.-K., Shum, H.-Y.: Drag-and-drop pasting. *ACM Siggraph* 25(3), 631–636 (2006)
3. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. *ACM Siggraph* 22(3), 313–318 (2003)

4. Farbman, Z., Hoffer, G., Lipman, Y., Cohen-Or, D., Fattal, R., Lischinski, D.: Coordinates for instant image cloning. *ACM Trans. on Graphics* 28(3), 67:1–67:10 (2009)
5. Chen, T., Cheng, M.-M., Tan, P., Shamir, A., Hu, S.-M.: Sketch2photo:Internet image montage. *ACM Trans. Graph.* 28(5) (2009)
6. Ding, M., Tong, R.: Content-aware copying and pasting in images. *The Visual Computer* 26, 721–729 (2010)
7. Sun, J., Jia, J., Tang, C.-K., Shum, H.-Y.: Poisson matting. *Proceedings of ACM SIGGRAPH* 23(3), 315–321 (2004)
8. Levin, A., Lischinski, D., Weiss, Y.: A closed form solution to natural image matting. In: *Proceedings of IEEE CVPR* (2006)
9. Rother, C., Kolmogorov, V., Blake, A.: Grabcut - interactive foreground extraction using iterated graph cut. *Proceedings of ACM SIGGRAPH* 23(3), 309–314 (2004)
10. Agarwala, A.: Efficient gradient-domain compositing using quadrees. *ACM Trans. Graph.* 26(3), 94:1–94:6 (2007)
11. Kazhdan, M., Hoppe, H.: Streaming multigrid for gradient-domain operations on large images. *ACM Trans. Graph.* 27(3), 21:1–21:10 (2008)
12. Bolz, J., Farmer, I., Grinspun, E., Schroder, P.: Sparse matrix solvers on the GPU: conjugate gradients and multigrid. *ACM Trans. Graph.* 22(3), 917–924 (2003)
13. Tao, M.W., Johnson, M.K., Paris, S.: Error-Tolerant Image Compositing. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010, Part I*. LNCS, vol. 6311, pp. 31–44. Springer, Heidelberg (2010)
14. Floater, M.S.: Mean value coordinates. *CAGD* 20(1), 19–27 (2003)
15. Hormann, K., Floater, M.S.: Mean value coordinates for arbitrary planar polygons. *ACM Transactions on Graphics* 25(4), 1424–1441 (2006)
16. Lipman, Y., Kopf, J., Cohen-Or, D., Levin, D.: GPU-assisted positive mean value coordinates for mesh deformations. In: *SGP 2007*, pp. 117–123 (2007)

Directional Eigentemplate Learning for Sparse Template Tracker

Hiroyuki Seto, Tomoyuki Taguchi, and Takeshi Shakunaga

Okayama University

Abstract. Automatic eigentemplate learning is discussed for a sparse template tracker. Using an eigentemplate learned from multiple sequences, a sparse template tracker can efficiently track a target that changes appearance. The present paper provides a feasible solution for eigentemplate learning when multiple image sequences are available. Two types of eigentemplates are compared in the present paper, namely, a single eigentemplate, and a set of directional eigentemplates. The single eigentemplate simply consists of all images learned from multiple sequences. On the other hand, directional eigentemplates are obtained by decomposing the single eigentemplate into three directions of the face poses. The sparse template tracker is also expanded to directional eigentemplates. Finally, the effectiveness of the provided solution is demonstrated in the learning and tracking experiments. The experimental results indicate that directional learning works well with small seed data, and that the directional eigentracker works better than the single eigentracker.

1 Introduction

Object tracking is one of the most significant problems in computer vision [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15]. Considerable research on unknown objects and known objects has been conducted in a lot of applications. Among them, some robust algorithms were proposed for the tracking based on the eigenspace techniques [5] with combining iterative projections and outlier detection. The iterative projection approaches, however, often suffer from time-consuming implementation and the “breakdown point” problems. In order to solve these problems, a sparse eigentemplate tracker was proposed by Shakunaga et al. [10] based on a non-eigentemplate tracker [8]. In the tracker proposed by Shakunaga et al., a particle filter is used in order to avoid iterative calculations. Shakunaga-Noguchi [16] demonstrated that the tracker [8] could be converted to an adaptive tracker by combining their sparse template tracking and an on-line learning technique of Black-Jepson [1]. Sakabe et al. [17] used the adaptive tracker for learning the eigentemplate used in the sparse eigentemplate tracking and demonstrated that the effectiveness of the learning. A tracker with memory-based particle filter was developed from the sparse template tracker [8] by Mikami et al. [12, 13]. In their tracker, past history storage is used for robust tracking by calculating the posterior face position with this storage. However, this tracker requires a large

amount of past data. In order to carry out robust eigentemplate tracking using a small amount of image data, the present research is based on the tracker proposed in [17].

In [17], an eigentemplate is learned from an image sequence. If the eigentemplate is learned from multiple sequences, then the tracker is expected to track more efficiently in case of changes in appearance. In the generation of eigentemplates, two types of eigentemplates, namely, a single eigentemplate and directional eigentemplates, are compared. Directional eigentemplates are obtained by classifying face templates into three directions. When the directional eigentemplates are used for eigentracker, the tracker selects appropriate eigentemplate in each frame and reduces the overmatching effect of the unified eigentemplate with respect to inappropriate poses. In addition, since the tracker evaluates the poses of the target with three eigentemplates, the tracker can avoid converging to a local minimum.

2 Learning Eigentemplate for Sparse Template Tracker

2.1 Adaptive Sparse Template Tracker

Automatic eigentemplate learning is possible, if an adaptive tracker is provided. If the tracker can carry out complete and accurate tracking for the case in which changes in appearance occur, then the problem of eigentemplate learning is reduced to a simple problem. However, since there is no such complete tracker, we must develop a learning method for a given tracker. The present paper basically uses an adaptive sparse template tracker formulated in Shakunaga-Noguchi [16]. The tracker is not complete but good since it combines the sparse template tracker and the WSL model proposed by Jepson et al. [1] for implementing an adaptive real-time tracker. In their formulation, the WSL model is applied to each pixel value, and an adaptive template, called the WSL template, is updated by the on-line EM algorithm. During the updating phase, an image estimated by the sparse template tracker is used to update the WSL model. Then, a dense template is constructed from the adaptive template, and the sparse template is updated. Thus, the tracker can carry out adaptive real-time tracking and sequential learning.

2.2 Learning Eigentemplate

This paper basically uses the eigentemplate learning formulated by Sakabe et al. [17]. Their method is summarized as follows: In tracking with the adaptive tracker, the estimated image is evaluated at each frame.

Their formulation uses the following notation. Let \mathbf{Y}_t and $\tilde{\Phi} = [\tilde{\Phi} \ \bar{\mathbf{x}}]$ denote an input image and the eigentemplate at time t , respectively. Let $Q_i (i = 1, 2, 3, 4)$ denote partial indicator matrices which correspond to four quadrant regions of the entire template, respectively, and let $Q_0 = Q_1 + Q_2 + Q_3 + Q_4 = I$ hold. Then, for $i = 0, 1, 2, 3, 4$, a projection of a (partial) image $Q_i \mathbf{Y}_t$ onto the

(homogeneous) eigentemplate, $\tilde{\Phi}$, is represented as $\mathbf{Y}'_{ti} = \tilde{\Phi}(Q_i\tilde{\Phi})^+\mathbf{Y}_t$. Thus, the correlation $C_i(\mathbf{Y}_t, \mathbf{Y}'_{ti})$ is calculated between $Q_i\mathbf{Y}_t$ and $Q_i\mathbf{Y}'_{ti}$.

Although Sakabe et al. [17] provided an image selection rule, a simpler rule is used in the present paper. That is, when the following condition is satisfied, the current input image \mathbf{Y}_t is appended to the learning set. Otherwise, the current image is not appended to the learning set.

$$\min_{i=0,1,2,3,4} C_i(\mathbf{Y}_t, \mathbf{Y}'_{ti}) < 0.7 \quad (1)$$

2.3 Sparse Eigentemplate Tracker

When an eigenspace is constructed from a set of normalized template images, it is used as an eigentemplate. The formulation of sparse template matching [16] can be generalized to eigentemplate matching as follows:

Let $\bar{\mathbf{x}}$ and Φ denote the mean vector and a matrix composed of the m most significant eigenvectors. Let $\tilde{\Phi}$ denote $[\Phi \ \bar{\mathbf{x}}]$. Then, the eigentemplate matching problem is formulated as follows:

$$\arg \min_{T \in \{T\}} \epsilon = \arg \min_{T \in \{T\}} \hat{\rho}\left(\frac{1}{\beta} P[\tilde{\Phi}\tilde{\mathbf{y}}^* - T\mathbf{Y}]\right), \quad (2)$$

where $\hat{\rho}(\mathbf{x})$ indicates the summation of the Geman-McClure function, $\tilde{\mathbf{y}}^*$ is an $(m+1)$ -vector calculated for each T as $\tilde{\mathbf{y}}^* = (P\tilde{\Phi})^+T\mathbf{Y}$, and β is a normalization parameter calculated for each T .

3 Learning and Tracking for Multiple Sequences

3.1 Eigentemplate Learning for Multiple Sequences

Once an eigentemplate is learned from an image sequence, the tracker can track similar sequences using the eigentemplate. If an eigentemplate consists of images learned from more varied image sequences, the tracker is expected to track against more varied changes of appearances. Therefore, in the present paper, two types of eigentemplates learning are considered for multiple sequences. In 3.2, simple expansion of Sakabe et al.'s method is discussed. The other expansion is discussed in 3.3, where a set of directional eigentemplates are learned from multiple sequences.

3.2 Simple Expansion of Sakabe et al.

As a simple expansion of the eigentemplate learning two types of learning should be considered for multiple sequences.

The first type is parallel learning, in which each sequence is first used to obtain a learning set of images independent of the other sequences. Then, the learning sets, selected from each sequence, are merged to generate a single eigentemplate.

Therefore, the result of parallel learning is invariant with respect to the order of image sequences. On the other hand, the result may be redundant since each parallel learning is carried out without any initial information.

The second type is cascade learning, which learns one sequence after another. Since cascade learning starts from the eigentemplate obtained from other sequences, only a small number of images are learned in each sequence. On the other hand, the results of learning may depend on the order of sequences used in learning. In the present paper, the parallel learning is used for multiple learning because no order problem is included in the learning.

3.3 Directional Eigentemplates

Once the single eigentemplate is learned from multiple sequences, the tracker is expected to track efficiently for all of the changes in appearance in multiple sequences. However, if the eigentemplate is constructed from too large a set of various images without considering the face poses, the tracker may excessively match inappropriate poses of the target. Actually, some combinations of multiple sequences often result in inefficient tracking. In such cases, the single eigentemplate causes unstable tracking when the pose estimation error is generated.

In order to avoid such a critical problem, we consider decomposing an eigentemplate into three directions of the face (front, left, right) as shown in Fig. 1. We call this set of eigentemplates “directional eigentemplates”. By decomposing the eigentemplate, the tracker is expected to select an appropriate eigentemplate for the poses of the target. Therefore, the tracker will reduce overmatching and avoid unstable tracking.

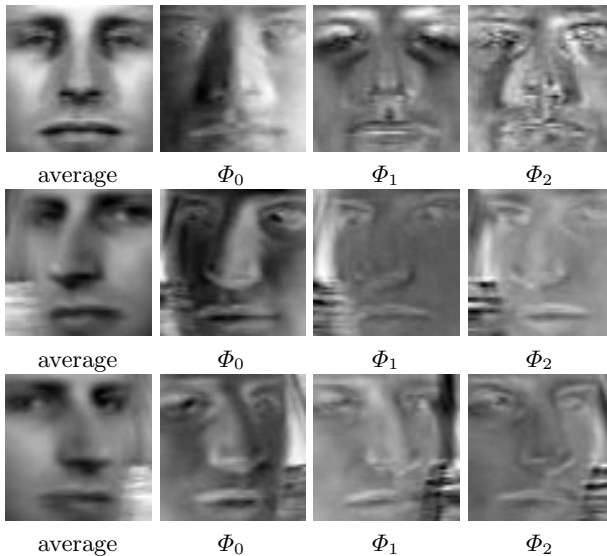


Fig. 1. Directional eigentemplates(front, left, right)

3.4 Automatic Learning of Directional Eigentemplates

Learning of directional eigentemplates consists of two parts, “direction judgment” and “learning judgment”. In order to discuss how to select the learning set, we use the following notation. Let $[\Phi_d, \bar{\mathbf{x}}_d]$ denote the seed directional eigentemplates of each direction, and let \mathbf{Y}_{0d} denote the initial image of each direction (d=f,l,r). The seed directional eigentemplates are provided in order to determine the direction of each frame. Each of the eigentemplates consists of a few images captured under different lighting conditions.

In the direction judgment, a projection \mathbf{Y}'_{td} of input \mathbf{Y}_t is made for each seed directional eigentemplate. Then, the correlation between the input and the projection, $C_0(\mathbf{Y}_t, \mathbf{Y}'_{td})$, is calculated in each direction d. The direction providing the highest correlation is determined as the direction of frame t.

After the direction judgment, the learning judgment is performed. Four correlation are calculated between each quadrant image \mathbf{Y}_{ti} (i=1,2,3,4) and the partial projection image \mathbf{Y}'_{ti} . Let the i-th correlation be denoted as $C_i(\mathbf{Y}_{ti}, \mathbf{Y}'_{ti})$. If the correlation satisfies the condition(1), then the current input image \mathbf{Y}_t is appended to the learning set. Otherwise, the current image is not appended to the learning set.

3.5 Expansion for Eigentemplate Tracking

Next, the tracker is expanded to directional eigentemplates. Let $\bar{\mathbf{x}}_i$ and Φ_i (i=f,l,r) denote the mean vector and a matrix composed of the m most significant eigenvectors for each direction. Let $\tilde{\Phi}_i$ denote $[\Phi_i \bar{\mathbf{x}}_i]$. The particle filter first evaluates each particle and then selects the optimal particle in each direction as follows:

$$\arg \min_{T \in \{T\}} \epsilon = \arg \min_{T \in \{T\}} \hat{\rho} \left(\frac{1}{\beta} P[\tilde{\Phi}_i \tilde{\mathbf{y}}^* - T\mathbf{Y}] \right), \quad (3)$$

Next, the proportion of each direction in the top particles are calculated by comparing ϵ , where top particles are a set of the best particles used to estimate the position of the next frame. Finally, the top particles are selected according to the proportion. In the selection phase, the top particles are basically selected from the direction that provides the highest proportion. When the highest proportion is less than 0.80, top particles are selected from the highest and the second directions. In this way, the tracker is expected to perform stable pose estimation as the pose of the target changes.

4 Experiments

4.1 Learning Directional Eigentemplates

Let us perform eigentemplate learning on the image sequences of Cascia et al. [7]. In this experiment, 30 trials of a set of directional learning were first carried out for each **jal** sequence. For directional learning, a set of directional images, as shown in Fig. 2, was provided for the seed directional eigentemplates. Since a



Fig. 2. Images for the seed directional eigentemplates. Three seed images are shown in each row. The front, left, and right directions are shown from top to bottom.

set of directional eigentemplates was constructed in each trial, a total of 30 sets of directional eigentemplates were constructed from each **jal** sequence.

After learning, each sequence was tracked by the sparse eigentracker with a set of learned eigentemplates. In tracking experiment, we used the eigentemplates learned from **jal3,4,5,6** and **jal9** because these sequences included appropriate changes in appearances. In other words, **jal3,4**, and **9** include up-and-down sequences, and **jal5** and **6** include right-and-left sequences. In some cases, the number of direc-

Table 1. Success rates(%) of learning and tracking with “s”ingle eigentemplate(S) and “d”irectional eigentemplates(D).(X) under “D” indicates direction(s) used for directional tracking(“f”ront(F),“l”eft(L),“r”ight(R)). Test sequences were tracked with the eigentemplate learned from each learning sequence.

test sequence	learning sequence									
	jal3		jal4		jal5		jal6		jal9	
	S	D (F)	S	D (F)	S	D (LFR)	S	D (LFR)	S	D (F)
jal1	30	17	91	93	74	97	72	50	61	70
jal2	28	13	87	73	100	100	99	100	100	100
jal3	46	43	81	90	8	47	3	57	84	53
jal4	71	70	97	97	0	60	0	57	73	33
jal5	0	0	2	3	100	100	85	43	81	57
jal6	0	7	2	10	83	90	100	100	33	33
jal7	51	20	53	77	100	100	100	100	99	97
jal8	22	7	85	87	100	100	100	100	93	90
jal9	8	10	88	83	0	100	0	100	95	93
average	28.4	20.7	65.1	66.7	62.8	87.8	62.1	78.5	79.9	69.6

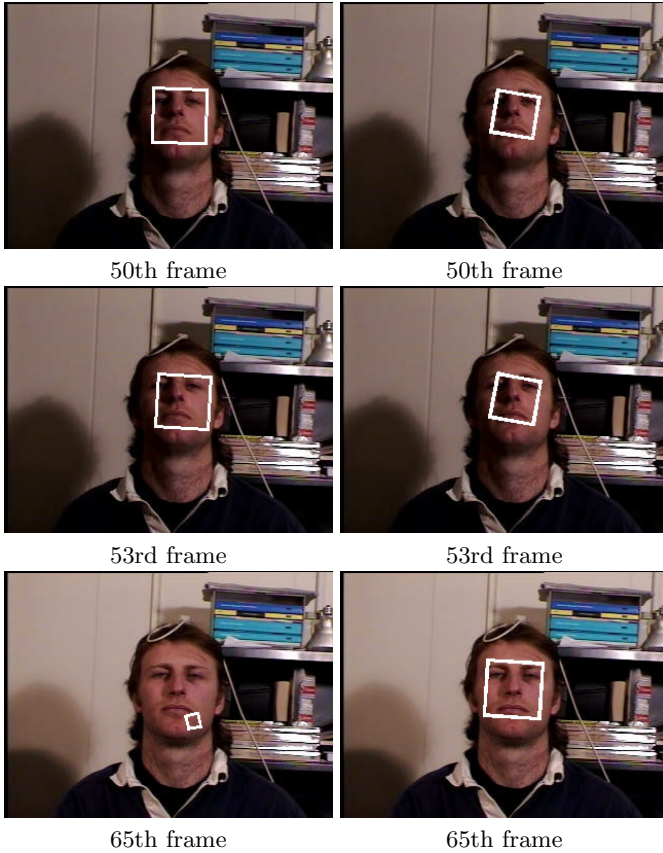


Fig. 3. Comparison of tracking *jal4* with two types of eigenfaces learned from *jal6*(left: tracking with a single eigentemplate, right: tracking with directional eigentemplates)

tional eigentemplates varies from the results of the learning (the directions used are listed in Table 1). After tracking, the results were compared with the correct data for each sequence. In the evaluation, the averages of the estimation errors in distance [pixel], rotation [deg], and scale, are used for evaluation. If the averages satisfy $\bar{d} < 4.0$ [pixel], $\bar{r} < 4.0$ [deg], and $\bar{s} < 0.2$, then the tracking is judged as “success”. Table 1 shows the success rates of the tracking with directional eigentemplates and those with single eigentemplate. The success rates indicate how well the eigentemplates that are effective for tracking are learned.

In the learning phase, for the most part, the images were correctly learned with respect to the directions. No confusion occurred between the right and left directions. In a few cases, images that might be regarded as front images were learned as left images. However, the eigentemplate could cover new appearances that the eigentemplate did not cover before the learning.

As shown in Table 1, the results of the directional eigentemplates were sometimes lower than those of single eigentemplates. However, the average success

Table 2. Success rates (%) of learning and tracking. The test sequences were tracked with the eigentemplates learned from each learning set. The configuration of this table is the same as that of Table 1.

test sequence	learning sets													
	jal3+jal5		jal4+jal6		jal4+jal5 +jal6		jal4+jal6 +jal9		jal3+jal4 +jal5+jal6		jal4+jal5 +jal6+jal9		jal3+jal4 +jal5+jal6 +jal9	
	S	D	S	D	S	D	S	D	S	D	S	D	S	D
jal1	40	40	100	80	90	83	90	77	77	87	93	87	83	90
jal2	53	67	100	97	100	100	100	100	87	100	100	100	100	100
jal3	60	40	87	83	57	67	80	97	83	73	50	50	87	83
jal4	77	70	100	97	100	93	100	100	100	100	100	83	100	100
jal5	100	97	0	20	90	100	0	27	63	100	93	100	63	100
jal6	37	93	100	100	77	100	100	100	67	100	87	100	63	100
jal7	67	63	100	97	100	100	100	100	97	100	100	100	100	100
jal8	67	73	100	97	100	100	100	100	87	100	100	100	100	100
jal9	67	70	100	97	100	100	100	100	93	100	100	100	100	100
average	63.0	67.8	87.4	85.2	90.0	93.7	85.6	88.9	83.7	95.6	91.5	91.1	88.5	97.0

rates were better with directional eigentemplates than those with a single eigentemplate. In particular, the averages of **jal5** and **jal6** increase considerably. The directional eigentemplates improved pose estimation by selecting an appropriate eigentemplate in each frame.

Examples of the tracking with two eigentemplates are as shown in Fig. 3. With a single eigentemplate, tracking was stable until the 50th frame. However, the pose estimation error occurred at the 53rd frame, and the error continued. Finally, the tracker converged to local minimum at the 65th frame. On the other hand, the pose estimation error occurred until the 53rd frame in the tracking with directional eigentemplates. However, the tracker gradually corrected the pose of the target, and the error was resolved at the 65th frame. The results show that the directional eigentemplates are efficient for the pose estimation error on the tracking and provide efficient tracking.

4.2 Learning Eigentemplates from Multiple Sequences

The results of the previous experiment revealed that an eigentemplate learned from an image sequence can track other sequences to a certain extent. However, the eigentemplate often fails to track certain sequences because the eigentemplate includes information included in the learning sequence. In the single eigentemplate tracking, the eigentemplate learned from **jal3** could not track **jal6**, whereas the eigentemplate learned from **jal5** could track **jal6**. Therefore, if the eigentemplate is learned from **jal3** and **jal5**, the tracker is expected to carry out stable tracking **jal6**. In the following experiment, we tried to construct the eigentemplate from multiple sequences.



Fig. 4. Comparison of tracking **jal6**(left:tracking with directional eigentemplate learned from **jal3**,right:tracking with directional eigentemplates learned from **jal3+jal5**)

In the experiment, we also compared the two types of eigentemplates using tracking **jal** sequences. After learning each sequence, the images learned from some sequences were combined to include the information of other sequences. Therefore, the learned images included different poses, such as **jal3 + jal5**. (Images learned from **jal3** include up-and-down information, and images learned from **jal5** include right-and-left information.) The combinations of the learned images were as shown in Table 2. The tracking was carried out 30 times for each sequence with each eigentemplate.

Table 2 compares the success rate of tracking using a single eigentemplate(S) and the directional eigentemplates(D). In the table, the results were evaluated similar to the manner described in 4.2.

The results of tracking are shown in Table 2, which indicates that the tracker can carry out stable tracking when the eigentemplate is learned from multiple sequences. For example, the results for tracking **jal6** were better with the

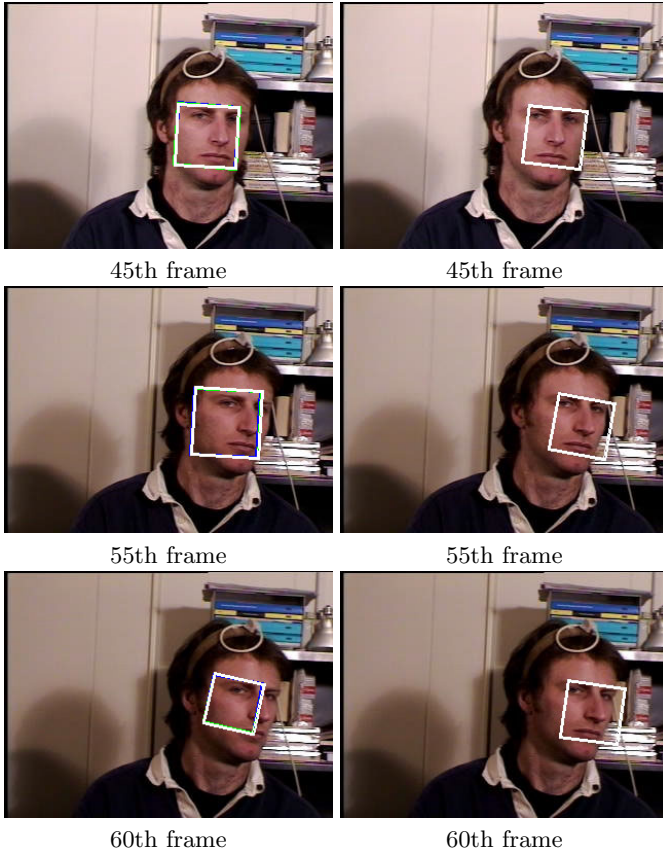


Fig. 5. Comparison of tracking **jal5** with two types of eigentemplates learned from **jal4+jal6+jal9**(left:tracking with single eigentemplate, right:tracking with directional eigentemplates)

eigentemplate learned from **jal3** and **jal5** than that learned from each sequence, as shown in Fig. 4. The eigentemplate adapted to new appearance changes. However, in some cases, the results became worse with additional eigentemplate learning. For example, although, **jal5** was tracked stably with the eigentemplate learned from **jal6**, the tracker perform out the stable tracking when the eigentemplate was learned from **jal4** and **jal6**. The images learned from **jal4** and **jal6** were inappropriate for tracking **jal5**, since the combinations of face positions and lighting conditions obtained from **jal4** and **jal6** were different from those of **jal5**. Therefore, the tracker could not estimate the correct position.

Comparing the results of tracking using two different types of eigentemplates, the tracking with directional eigentemplates worked better than that with a single eigentemplate. In some case, results of directional eigentemplates were lower than those of a single eigentemplate. However, the average success rates was higher than a single eigentemplate for most combinations, which indicates that

the directional eigentemplates could use the information included in multiple sequences more effectively than the single eigentemplate.

The example shown in Fig. 5 indicates how the tracker worked using the two different types of eigentemplates. With single eigentemplate, a pose estimation error occurred at the 45th frame, after which the error continued. Finally, the tracker converged at a local minimum at the 60th frame. In contrast, for the same sequence, the tracker using directional eigentemplates could track the target correctly. In the sequence, when the target faced toward the right, the tracker selected the right eigentemplate. Therefore, the tracker could efficiently estimate the appropriate poses and track using the directional eigentemplates.

5 Conclusion

Directional eigentemplate learning was discussed for a sparse template tracker. In the learning phase, the adaptive tracker adaptively tracks a target for the eigentemplate learning. If an eigentemplate is decomposed into directional eigentemplates, then the sparse eigentemplate tracker can estimate the pose of the target with an appropriate eigentemplate.

The experimental results show that the directional learning worked well using a few initial images, and the tracking worked well using directional eigentemplates learned from single image sequences. In the second experiment, the tracker with directional eigentemplates was shown to work better than the single eigentemplate for multiple learning. In some cases, however, the tracker did not work well. In the future, we would like to solve the problems involved in these cases and develop a more stable on-line learning method.

This work has been supported in part by a Grant-In-Aid for Scientific Research (No.20300067) from the Ministry of Education, Science, Sports, and Culture of Japan.

References

1. Jepson, A.D., Fleet, D.J., El-Maraghi, T.F.: Robust online appearance models for visual tracking. *IEEE Trans. Pattern Analysis and Machine Intelligence* 25(10), 1296–1311 (2003)
2. Isard, M., Blake, A.: Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision* 29(1), 5–28 (1998)
3. Williams, O., Blake, A., Cipolla, R.: A sparse probabilistic learning algorithm for real-time tracking. In: *Proc. ICCV*, pp. 353–360 (2003)
4. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence* 24(5), 603–619 (2002)
5. Black, M., Jepson, A.: Eigentracking: Robust matching and tracking of articulated objects using a view-based representation. *International Journal of Computer Vision* 26(1), 63–84 (1998)
6. Avidan, S.: Ensemble tracking. *IEEE Trans. Pattern Analysis and Machine Intelligence* 29(2), 261–271 (2007)

7. Cascia, M.L., Sclaroff, S., Athitsos, V.: Fast, reliable head tracking under varying illumination: An approach based on robust registration of texture-mapped 3d models. *IEEE Trans. Pattern Analysis and Machine Intelligence* 22(4), 322–336 (2000)
8. Matsubara, Y., Shakunaga, T.: Sparse template matching and its application to real-time object tracking. *IPSJ Transactions on Computer Vision and Image Media* 46, 60–71 (2005) (in japanese no.sig9(cvim11))
9. Satake, J., Shakunaga, T.: Multiple target tracking by appearance-based condensation tracker using structure information. In: *Proc. International Conference on Production Research*, vol. 3, pp. 294–297 (2004)
10. Shakunaga, T., Matsubara, Y., Noguchi, K.: Appearance tracker based on sparse eigentemplate. In: *Proc. Int'l Conf. on Machine Vision & Applications*, pp. 13–17 (2005)
11. Oka, Y., Kuroda, T., Migita, T., Shakunaga, T.: Tracking 3d pose of rigid object by sparse template matching. In: *Proc. the 5th International Conference on Image and Graphics, ICIG 2009*, pp. 390–397 (2009)
12. Mikami, D., Otsuka, K., Yamato, J.: Memory-based particle filter for face pose tracking robust under complex dynamics. In: *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pp. 999–1006 (2009)
13. Mikami, D., Otsuka, K., Yamato, J.: Memory-Based Particle Filter for Tracking Objects with Large Variation in Pose and Appearance. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010, Part III. LNCS*, vol. 6313, pp. 215–228. Springer, Heidelberg (2010)
14. Murphy-Chutorian, E., Trivedi, M.M.: Hyhope: Hybrid head orientation and position estimation for vision-based driver head tracking. *IEEE Trans. Pattern Analysis and Machine Intelligence* 31(4), 607–626 (2009)
15. Oka, Y., Shakunaga, T.: Sparse eigentracker augmented by associative mapping to 3d shape. In: *Proc. IEEE Conference on Automatic Face and Gesture Recognition (FG 2011)*, pp. 649–656 (2011)
16. Shakunaga, T., Noguchi, K.: Robust tracking of appearance by sparse template adaptation. In: *Proc. 8th IASTED Int'l Conf. on Signal and Image Processing*, pp. 85–90 (2006)
17. Sakabe, K., Taguchi, T., Shakunaga, T.: Automatic Eigentemplate Learning for Sparse Template Tracker. In: Wada, T., Huang, F., Lin, S. (eds.) *PSIVT 2009. LNCS*, vol. 5414, pp. 714–725. Springer, Heidelberg (2009)

Gender Identification Using Feature Patch-Based Bayesian Classifier

Shen-Ju Lin, Chung-Lin Huang, and Shih-Chung Hsu

Department of Electrical Engineering
National Tsing-Hua University, Hsin-Chu, Taiwan, R.O.C.
g9761531@oz.nthu.edu.tw, clhuang@ee.nthu.edu.tw,
chvjohnff@gmail.com

Abstract. In the paper, we propose a Bayesian classifier which exploits non-parametric model to identify the gender from the facial images. Our major contribution is that we use feature patch-based non-parametric method to generate the posteriori of male and female based on the characteristics of the labeled training image patches. Our system consists of four modules. First, we use AAM model to identify facial feature points. Facial images are represented by the overlapping feature patches around the feature points. Second, from the labeled training patches, we select a smaller subset as the patch library based on the K means clustering. Third, in training, we embed the gender characteristics of the training feature patches as the posteriori of the library patches. Fourth, in testing, we integrate the posterior of the test patches to determine the gender. The experimental results demonstrate that our proposed method is better than the conventional non-feature-patch-based methods.

Keywords: Gender identification, Active Appearance Model, Patches-based Bayesian estimation.

1 Introduction

Biometric features of human faces reveal lots of high-level semantic information of the human such as gender, age, ethnicity and emotion expression and etc. Compared with age or ethnicity estimation, determining the gender of a facial image has become an interesting research topic. How human being identifying the gender is unknown and gender misjudgment often occurs. The facial images may contain the variation in illumination, pose, background clutter, and partial occlusion. We consider all these variations in facial image and develop a reliable method to identify the gender.

Gender identification methods can be divided into two main categories: geometry-based and appearance-based. The geometry-based category focuses on extracting the geometric feature points from facial images and describing the shape structure of the face. It uses Active Appearance Model (AAM) to build statistical model of object shape and texture information. Saatci *et al.* [1] propose an approach to determine the gender and expression of facial images by using AAM for feature extraction and Support Vector Machines (SVMs) for classification. Mäkinen *et al.* [2] present a systematic evaluation on gender classification, and show that how did face alignment influence the accuracy of gender classification methods.

The appearance-based category can further be divided into two approaches: texture-based and statistical-based. The former uses different texture descriptors to characterize a facial image about gender, and employ machine learning algorithm to classify the gender. The texture are modeled by local binary pattern (*LBP*) [3, 5], Local Gabor binary mapping pattern (*LGBMP*) [4], wavelet transform [6, 7], Ada-boost [8, 9], Independent Component analysis (*ICA*) [10]. Other researches [11] stress on the specific local region, such as nose, eyes, mouth, etc.

The statistical-based approach aims at using different features which are quantified into probability to characterize a facial image about gender according to their visual traits. Toews *et al.* [12] present the combination of local scale-invariant features (*SIFT*) and object class invariant (*OCI*) model for detecting, localizing and classifying visual traits of gender from facial images. Aghajanian *et al.* [13] propose a patch-based framework to determine the ambiguous within-object and replace each patch from the predefined library and frequency parameters corresponding to these patches. Li *et al.* [14] provide another patch-based feature representation called Spatial Gaussian Mixture Models (*SGMM*) to describe the image spatial information relative precisely at both local and global scales for image.

Different from [13, 14], we use non-parametric statistical method to embed the gender characteristics of the feature patches in a pre-defined patch library. In training process, we propose a Bayesian statistical framework to model the gender characteristics of the training facial images and build a posteriori gender probability distribution of the library patches. In the testing process, we integrate the gender posteriori distribution of the test image patches based on the patch library to predict the gender.

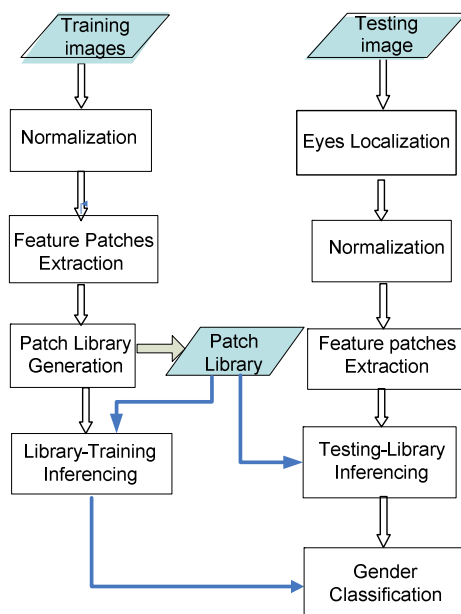


Fig. 1. The flowchart of gender estimation system

First, we apply Adaboost algorithm [8, 9] to find the locations of eyes and then normalize the face according to the center positions of both eyes. Then, we use AAM to locate the facial feature points and then extract the overlapping feature patches of facial image. Our approach consists of four modules: (1) Select a predefined patch library from the training image patches. (2) Generate the posterior of each feature patch in the library as the off-line library-training inference. (3) Generate the posterior of each feature patch in the input test image as the on-line testing-library inference. (4) Marginalize over all the feature patches to determine the classification decision. The contributions in this paper are applying (a) multiple overlapping feature patches, (b) Bayesian gender determination framework based in the off-line library-training inference and on-line test-library inference, and (c) a library selection scheme based on eigenface with K means clustering.

2 Facial Feature Points Extraction Using AAM

Based on the MORPHY database, we may normalize and rotate the face database and apply the AAM [15] to represent the face and extract the facial feature points. Because of the different angle and size of the face, it is necessary to normalize the face image for gender estimation. Since the eyes are easily found compared to the other face features, we normalize the face images based on both eyes to adjust the orientation and the size of the face. Then we rotate the line linking two centers and scale the distance between both eyes to normalize the region size of facial features. Finally, the hair region at the top of the image is also excluded.

We apply the AAM to an input image to find the model parameters by maximizing the “match” between the model instance and the input image. Then, the model parameters are used to find the facial feature points. AAM is a well-known statistical model which consists of two parts: the shape model and the texture model. To train the AAM model, we have a set of landmark points selected as the salient points on the human face as shown in Figure 2. For each image, we use the pre-trained AAM model to search for the 27 facial feature points which are located at the corner and bag regions of both eyes, the left and right jaw grain regions and the hair region.

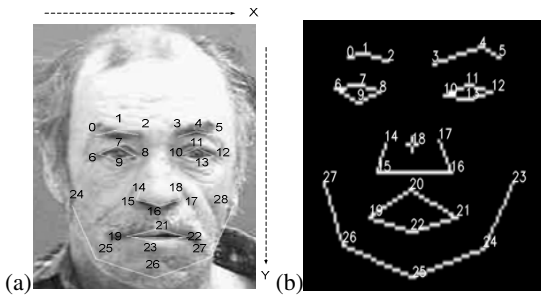


Fig. 2. (a) The label feature points of the training set. (b)The position of 28 landmark points.

3 Feature Patch-Based Gender Classification

Each face image is represented by a grid of overlapping feature patches. Based on the 27 facial feature points identified by AAM, there are 27 feature patches extracted as shown in Figure 3. Each feature patch is an independent individual that provides gender information. The gender classification consists of two processes: *library-to-training inference* and *testing-to-library inference*. The first process is an off-line training process, in which we build a non-parametric statistics model to embed the training images gender characteristics into the predefined library patches. In the on-line gender identification process, the posteriori of each test image patches is modeled by patches from the predefined library. By integrating the posterior probability of each feature patch, we may determine the gender of the face image.

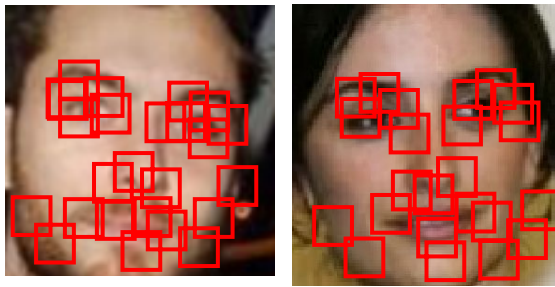


Fig. 3. The 27 selected feature patches

Gender is viewed as a class label C , where $C=Male$ or $C=Female$. The test image \mathbf{Y} is represented as a regular grid of overlapping feature patches $\mathbf{Y} = [y_1, y_2, \dots, y_N]$, where N is the number of selected feature patches. The resolution of the patch is $M \times M$. We choose a set of labeled training images from each class (male or female) to train the model. Each training image is represented by a regular grid of overlapping patches of the same size. In addition, we choose the images evenly from each of the two classes to build the predefined library, which are different from the training images. The predefined library consists of a set of overlapping feature patches which are of the same size as those in the training and testing images.

We consider the Bayesian estimation approach to identify the gender of the facial images. The idea of Bayesian estimation is that we treat each patch as a random variable and using the training images to generate a posteriori probability density of the patch. The main goal of Bayesian classification lies at the computation of the posterior probability of facial image \mathbf{Y} , $P(C|\mathbf{Y})$, which can be derived based on the prior $P(C)$ and the class-conditional densities $P(\mathbf{Y}|C)$. Based on the set of training set \mathbf{X} , we may have the Bayesian posterior probabilities $P(C|\mathbf{Y}, \mathbf{X})$. From each training image, we may extract N ($=27$) overlapping feature patches. So, the training set \mathbf{X} can be represented by N patch groups as $\{\mathbf{X}^p | p=1, \dots, N\}$. Each patch group has two classes, such as $\mathbf{X}^p = \mathbf{X}^p_{male} \cup \mathbf{X}^p_{female}$. Here, we denote the p th feature patch of the i th training image as $\mathbf{x}_{pi} \in \mathbf{X}^p_c$, where c indicates either *male* or *female*.

Given the training set \mathbf{X} , we can obtain the posteriori probability of a facial image \mathbf{Y} of class C by using Bayes' rule as

$$P(C | \mathbf{Y}, \mathbf{X}) = P(\mathbf{Y} | C, \mathbf{X}) \cdot P(C) / P(\mathbf{Y}) \propto \prod_{p=1}^N P(\mathbf{y}_p | C, \mathbf{X}) \cdot P(C) \quad (1)$$

where \mathbf{y}_p is an individual test patch of the facial image \mathbf{Y} . Based on equation (1), we may find the posteriori $P(C | \mathbf{Y}, \mathbf{X})$, and the gender of the image \mathbf{Y} can be determined by maximizing a posteriori as

$$\text{Gender}(\mathbf{Y}) = \text{Argmax}_c P(C | \mathbf{Y}, \mathbf{X}) \quad (2)$$

Here, we have constructed the library with 27 library subsets for each class as M^c_1, \dots, M^c_{27} , of which each has L feature patches as $M^c_p = \{\mu_{cpl} | l=1, \dots, L\}$. $P(\mathbf{y}_p | \mu_{cpl})$ with the relational information of parameter space and training set can be used to predict the gender probability of \mathbf{y}_p for $p=1, \dots, N$. To obtain the class-conditional densities $P(\mathbf{y}_p | C, \mathbf{X})$, we do the integrating of the joint density $P(\mathbf{y}_p, \mu_{cpl} | C, \mathbf{X})$ over a set of parameter $\{\mu_{pl}\}$ as

$$\begin{aligned} P(\mathbf{y}_p | C, \mathbf{X}) &= \int P(\mathbf{y}_p, \mu_{cpl} | C, \mathbf{X}) d\mu_{cpl} \\ &= \int P(\mathbf{y}_p | \mu_{cpl}) \cdot P(\mu_{cpl} | C, \mathbf{X}) d\mu_{cpl}. \end{aligned} \quad (3)$$

$\{\mu_{cpl}\}$ denotes a set of library patches, of which the indices p and l indicate the p^{th} feature patch in the l^{th} library image. In Eq. (3), the class-conditional densities $P(\mathbf{y}_p | C, \mathbf{X})$ consists of a chained probability of $P(\mathbf{y}_p | \mu_{cpl})$ and $P(\mu_{cpl} | C, \mathbf{X})$ where $P(\mu_{cpl} | C, \mathbf{X})$ can be viewed as embedding the gender characteristics of the training feature patches in \mathbf{X} into the patch library space $\{\mu_{cpl}\}$ for $c=1$ or 2 , $1 \leq i \leq I$, $1 \leq p \leq 27$, and $1 \leq l \leq L$.

For each test patch \mathbf{y}_p , we have two cascaded terms $P(\mathbf{y}_p | \mu_{cpl})$ and $P(\mu_{cpl} | C, \mathbf{X})$ which are modeled by two inference processes: testing-library inference and library-training inference. To determine the gender of the input image, we (1) cascade the two inferences, (2) integrate all patches, (3) multiple the gender likelihood of all the patches of the test image as $\prod_{p=1}^N P(\mathbf{y}_p | C, \mathbf{X})$, and (4) find the maximum posterior of $P(C | \mathbf{Y}, \mathbf{X})$ to determine the gender.

3.1 Patch Similarity Measure

Each pixel in the feature patch can be described by a set of local binary pattern (*LBP*) codes [3, 5]. *LBP* code encodes the local structure around each pixel in the feature patch. The histogram of *LBP* labels calculated over the patches can be exploited as a texture descriptor. A popular measure between two feature patches in terms of two normalized histogram of *LBP* labels $p(u)$ and $q(u)$ is the Bhattacharyya coefficient. For discrete densities such as the normalized *LBP* histograms $p(u) = \{p^{(u)}\}_{u=1 \dots m}$ and $q(u) = \{q^{(u)}\}_{u=1 \dots m}$, the correlation coefficient is defined as

$$\rho[p, q] = \sum_{u=1}^m \sqrt{p^{(u)} q^{(u)}} \quad (4)$$

The larger ρ is, the more similar these two feature patches are. The similarity between two patches can be defined as

$$d = \sqrt{1 - \rho[p, q]}. \quad (5)$$

Dissimilar distributions result in a larger d . Furthermore, the likelihood can be written as

$$p(x|\mu_l) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{d^2}{2\sigma^2}} \quad (6)$$

The likelihood is specified by a Gaussian with variance σ . So, we may have probability density distribution as

$$P(\mathbf{x}|\mu_{cpl}) = P(\mathbf{x}|\mu_{cpl}) / \sum_{l=1}^L P(\mathbf{x}|\mu_{cpl}) \quad (7)$$

where $P(\mathbf{x}|\mu_{cpl})$ is a probability density distribution with L bins for $\mu_{cpl}, \mu_{c2}, \dots, \mu_{cpl}$.

3.2 Training Process

The training set X consists of two labeled training images sets as $X = X_{male} \cup X_{female}$ which are further decomposed into N labeled patch groups as $X_{male} = \{X_{male}^p | p=1, \dots, N\}$ and $X_{female} = \{X_{female}^p | p=1, \dots, N\}$. We may also describe the two labeled training sets as $X_{male} = \{x_{cpi} | C=1, p=1, \dots, N, i=1, \dots, I\}$ and $X_{female} = \{x_{cpi} | C=2, p=1, \dots, N, i=1, \dots, I\}$. The library can be further divided into subgroup M_p^c , where the indices c and p indicate the p^{th} subset of class C .

Before calculate $P(\mu_{cpl} | C, X)$, we use Bayesian rule to derive the posteriori density $P(\mu_{cpl} | C, \mathbf{x}_{cpi})$ over the parameters space $\{\mu_{cpl}\}$. We simplify $P(\mu_{cpl} | C, \mathbf{x}_{cpi}) = P(\mu_{cpl} | \mathbf{x}_{cpi})$ for specific class c which is defined as

$$P(\mu_{cpl} | \mathbf{x}_{cpi}) = \frac{P(\mathbf{x}_{cpi} | \mu_{cpl}) P(\mu_{cpl})}{P(\mathbf{x}_{cpi})} \quad (8)$$

where \mathbf{x}_{cpi} denotes the p^{th} feature patch in the i^{th} training image of specific class c . The prior $P(\mu_{cpl})$ indicates the weight of the l th patch in the library subset M_p^c which is determined by the frequency of the designated patches being selected in the training process. $P(\mu_{cpl})$ is normalized by $P(\mu_{cpl}) / \sum_l P(\mu_{cpl})$. Then, we have

$$P(\mu_{cpl} | C=c, X) = \int P(\mu_{cpl} | C, \mathbf{x}_{cpi}) d\mathbf{x}_{cpi} / \sum_l P(\mu_{cpl} | C, X) \quad (9)$$

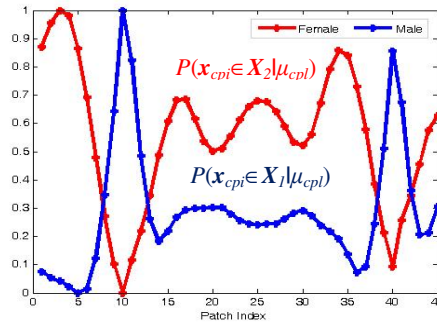


Fig. 4. The likelihood density functions $P(\mathbf{x}_{cpi} \in X_1 | \mu_{cpl})$ and $P(\mathbf{x}_{cpi} \in X_2 | \mu_{cpl})$ for every library patch μ_{cpl} , where $X_1 = X_{male}$ and $X_2 = X_{female}$.

To find the posteriori density $P(\mu_{cpl} | C, \mathbf{X})$ for class c , we calculate the likelihood term $P(\mathbf{x}_{cpi} | \mu_{cpl})$ of the labeled training sets \mathbf{X}_{male} and \mathbf{X}_{female} over the pertinent parameters μ_{cpl} , and exploit the relationship of the corresponding feature patch similarity between \mathbf{x}_{cpi} and μ_{cpl} . For each μ_{cpl} , we have the likelihood function $P(\mathbf{x}_{cpi} | \mu_{cpl})$ for each library patch μ_{cpl} as shown in Figure 4.

We use the *maximum a posteriori estimator* (MAP) to find the highest similarity between the training feature patch \mathbf{x}_{cpi} and all the possible library feature patch μ_{cpl} in M_p^c . We search all μ_{cpl} in M_p^c for the highest similar one with the training feature patch. For every training feature path \mathbf{x}_{cpi} , we find the most similar library feature patch in M_p^c as

$$\mu_{cpl}^* = \mathop{\text{Argmin}}_t \text{Dis}(\mu_{cpl}, \mathbf{x}_{cpi}) \text{ for each } \mathbf{x}_{cpi} \quad (10)$$

where $\mu_{cpl} \in M_p^c$, and M_p^c is the p th library subset, and $\text{Dis}(\mu_{cpl}, \mathbf{x}_{cpi})$ is calculated by using eq. (4).

To find the likelihood function $P(\mathbf{x}_{cpi} | \mu_{cpl})$ for specific library patch, we search all training feature patches to find the similarity. Considering I independent training feature patches \mathbf{x}_{cpi} , $i=1, \dots, I$, we can represent the likelihood of the training patches, $P(\mathbf{x}_{cpi} | \mu_{cpl})$, related to the specific parameter μ_{cpl} . With $P(\mathbf{x}_{cpi} | \mu_{cpl})$, we may obtain the posteriori distribution for the parameter set $\{\mu_{cpl}\}$ as

$$P(\mu_{cpl} | \mathbf{x}_{cpi}) \propto P(\mathbf{x}_{cpi} | \mu_{cpl}) \cdot P(\mu_{cpl}) / P(\mathbf{x}_{cpi}) \quad (11)$$

where $P(\mu_{cpl})$ is the prior of the parameter μ_{cpl} . The training feature patches correlated with predefined library feature patch are used to construct the posterior $P(\mu_{cpl} | \mathbf{x}_{cpi})$.

3.3 Testing Process

To predict the probability of the class of the input feature patch \mathbf{y}_p , we need to compute the likelihood $P(\mathbf{y}_p | \mu_{cpl})$ by exploiting the feature patch similarity between \mathbf{y}_p and μ_{cpl} . Then, we may combine the probability $P(\mathbf{y}_p | \mu_{cpl})$ and $P(\mu_{cpl} | \mathbf{X})$ to obtain the joint conditional density, $P(\mathbf{y}_p, \mu_{cpl} | C=c, \mathbf{X})$. For each feature patch \mathbf{y}_p , we may find its similarity with the library feature patch μ_{cpl} described as

$$P(\mu_{cpl} | \mathbf{y}_p) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{d^2}{2\sigma^2}} \quad (12)$$

where d is the similarity measure between two feature patches (Equ. (5)). We use the Bayes' rule to calculate the posteriori density $P(\mathbf{y}_p | \mu_{cpl})$ over the parameters $\{\mu_{cpl}\}$ as

$$P(\mathbf{y}_p | \mu_{cpl}) = \frac{P(\mu_{cpl} | \mathbf{y}_p) P(\mathbf{y}_p)}{P(\mu_{cpl})} \quad (13)$$

where prior $P(\mathbf{y}_p)$ indicates the weight of the p th test patch which is determined by the frequency of the designated patches being selected in the testing-library inference process.

To find the posteriori density $P(\mathbf{y}_p | \mu_{cpl})$, we calculate the likelihood term $P(\mu_{cpl} | \mathbf{y}_p)$ based on the subset M_p^c by exploiting the patches similarity measure (eq. (12)). We use the *maximum a posteriori estimator* (MAP) to find the highest similar feature

patch y_p for all the possible library feature patch μ_{cpl} in M_p^c . For every μ_{cpl} , we find y_p selected if

$$Dis(y_p, \mu_{cpl}) > \theta_p \quad (14)$$

where $Dis(y_p, \mu_{cpl})$ is calculated by using eq. (4), and θ_p is the similarity threshold. For each μ_{cpl} , we use (14) to determine whether the specific y_p is selected or not, and then accumulate the number of each y_p being selected for all μ_{cpl} . We divide the number by the total accumulated number of the all y_p as the priori $P(y_p)$, so that $\sum_p P(y_p) = 1$. With likelihood term $P(\mu_{cpl}|y_p)$ and $P(y_p)$, we may have the posteriori density $P(y_p|\mu_{cpl})$.

We construct the male and female posteriori distribution by using on the male and female training images. During the identification process, for every test patch, we generate its posterior information by search all library patches based on their similarity. Finally, we can integrate overall those patches from the facial image to provide a posterior probability and then determine the gender.

The male and female accumulated distribution for the same library image is obviously different, the extent of accumulated distribution corresponding to the gender can illustrate that the possibility of the local region in the library image belongs to the specific gender. The difference of the male and female specific accumulated distribution can represent the corresponding specific region in the library image having discriminating ability about gender.

4 Library Selection Scheme

The formation of predefined library is a crucial step for gender identification. In the training process, we calculate the accumulated frequency of all possible patches within the predefined library. In the inference process, we exploit the accumulated information of library patch which contain the highest similarity with the corresponding test patch. Basically, the library feature patches are obtained from the facial images of various kinds of changes such as in illumination, pose, and background clutter. We find the relation of the variation of predefined library composition with regard to the classification accuracy.

We propose a method by using the characteristic of eigenface with K-mean clustering for library selection. The eigenface method is based on principle component analysis (PCA), used to find a suitable low-dimensional space. There are two primary procedures: (a) **Eigenspace generation**. Given I normalized facial image patches for training, and each feature patch is represented as $M \times M$. These feature patches are converted into the column vector type with the dimension $M^2 \times 1$. (b) **Projection onto eigenspace**. Each of the training image patch x_i is projected onto the eigenspace as the weighting vectors. Weighting vector is calculated by the eigenspace and the average patch vector with inner product. The weighting vectors are considered as the feature points in K -dimension space.

With the weighting vectors from these patches, we execute the cluster analysis to group the weighting vectors by the variety of characteristics. Here, we choose K -means clustering to partition I patches into k clusters in which each patch is associated one of the clusters with the nearest mean, and the weighting vectors is viewed as the observations. After K -mean clustering, we choose each cluster centroid and the fea-

ture point with the minimum Euclidean distance as the selected patch. Finally, we collect these selected patches as a library.

5 Experimental Results

Here, we evaluate our gender classification algorithm by using different formation of library images. As shown in Figures 5.1~5.3, our face images are obtained from: (1) Wild database contains more than 13000 face images collected from 5749 people. (2) Caltech 10,000 web faces database contains 10524 human faces with various resolution and settings. (3) Color FERET database. In our experiment, we only use the *fa* partition in FERET database, *i.e.*, the regular frontal facial images. It contains 1364 images.



Fig. 5.1. Five male/female facial image in Labeled Faces in the Wild database



Fig. 5.2. Five male/female facial image in Caltech 10,000 web faces database

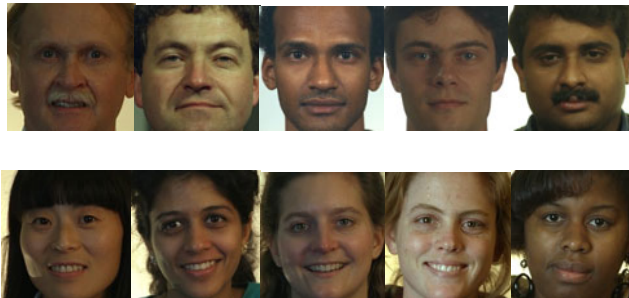


Fig. 5.3. Five male/female facial images in Color FERET database

From the male and female image patches, we select some patches as the library and let the remaining as the training image patches. We illustrate the difference of accurate gender identification according to the different library formation. To describe the patch information precisely, we set the size of patch to 6×6 grid block uniformly. Then, we select the testing set which consist 500 male and female images with various kinds of variations (lighting, expression, pose, background...etc). In the following, we show four different experiments and compare the performance of our method with the non-feature-patch-based method of which the library images are selected randomly.

(1) **Experiment 1.** The training set consists of 8000 male and 8000 female images to build the model. Then, we regulate the test images with two different resolution (30×30 , 60×60) and 7 different library formations ($L=2, 10, 30, 60, 120, 240, 360$) to show the difference of correct gender prediction. We compare the recognition rate of our method with [13] as illustrated in Figure 6 to show that our method are better than the random selection method in most cases. Our method uses more discriminative male and female patches so that the accuracy rate of our method is better. Smaller L indicates less information in the library, and the detection accuracy rate is lower. With more library patches, we have higher detection rate. By using the discriminative library patches with variation, our method shows better performance.

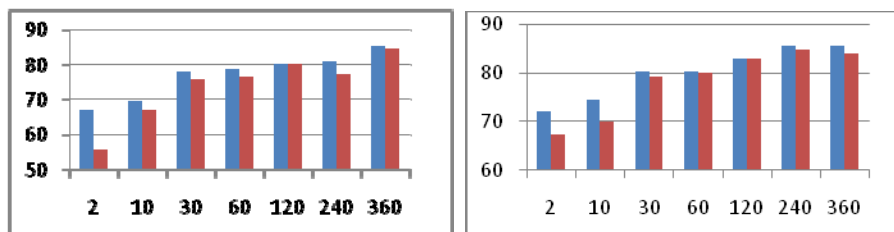


Fig. 6. The accuracy comparison of our method and [13] using (a) 30×30 test images. (b) 60×60 test images. (Training set: 8000 male and female images).

(2) **Experiment 2.** The smaller training set consists of 4000 male and 4000 female images to build the model, and select 7 different library formations ($L=2, 10, 30, 60, 120, 240, 360$). We also let the test images with two different resolution (30×30 , 60×60) and compare the gender identification rates of our method with [13] as illustrated in Figure 7. The experimental results show that the selection of library and training images from the smaller image database can still maintain a good gender classification performance. However, smaller L indicates less information in the library, and the detection rate is lower. Figure 7(b) shows that, for $L=120$, gender classification accuracy of our method and [13] are 85.5% and 82.5%, respectively.

(3) **Experiment 3.** We use the *fa* partition from color FERET database to verify our gender identification algorithm. The frontal facial images consist of 861 males and 503 females. Experiment 2 demonstrates that using clustering library selection with 240 library patches with the resolution 60×60 retains the best accurate gender identification rate. Using library with lots of image patches will provide more gender accumulated distribution information. The library does not contain enough characteristics of FERET

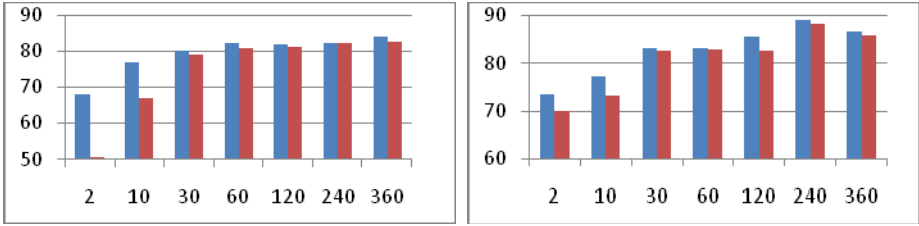


Fig. 7. The accuracy comparison of our method and [13] using (a) 30×30 test images. (b) 60×60 test images. (Training set: 4000 male and female images).

frontal facial images, so that, we have a less accurate gender prediction for FERET frontal facial images. The recognition accuracy of our method is 78.01% compared with 76.39% of the method [13].

(4) Experiment 4. We use Bao face database images for gender classification experiment from and show some gender detection results in Figures 8. We use two different library selection methods with $L=240$, and each face image with the resolution 60×60 , respectively. Based on the experimental results shown in Figure 8, we find that our method is better than [13]. The experimental results of different library formations are summarized and shown in Tables 1 and 2.



Fig. 8. Some gender detection result of the images use by the model with eigenface with clustering selection (Top) and random selection (Bottom)

Table 1. The recognition rate of two methods with different library $L=240/L=120$ with resolution 60×60

Recognition rate	Our method	Random selection[13]
$L=240$	88.7%	86.8%
$L=120$	85.5%	82.5%

Table 2. The accuracy comparison with library $L=8/L=20$, and image resolution 60×60

Recognition rate	Our method	Random selection[13]
$L=8$	83.4%	80.03%
$L=20$	72.9%	72.1%

6 Conclusions

We have proposed a modified Bayesian estimation framework to exploit patch similarity and accumulated distribution to predict human gender of the facial images. We propose a library selection scheme to choose the discriminative male and female images based on K means clustering, and build the male and female accumulated distribution based on the characteristics of the labeled training images. The experimental results demonstrate that our proposed method is better than the conventional method.

References

- [1] Saatci, Y., Town, C.: Cascaded Classification of Gender and Facial Expression Using Active Appearance Models. In: AFGR 2006, pp. 393–400 (April 2006)
- [2] Mäkinen, E., Raisamo, R.: Evaluation of Gender Classification Methods with Automatically Detected and Aligned Faces. *IEEE Trans. on PAMI* 30(3), 541–547 (2008)
- [3] Lian, H.C., Lu, B.L.: Multi-View Gender Classification Using Multi-Resolution Local Binary Patterns and Support Vector Machines. *Int. J. of Neural Systems* 17(6) (2007)
- [4] Xia, B., Sun, H., Lu, B.L.: Multi-View Gender Classification Based on Local Gabor Binary Mapping Pattern and Support Vector Machines. In: *IEEE ICNN*, p. 3388 (2008)
- [5] Fang, Y., Wang, Z.: Improving LBP Features For Gender Classification. In: *Int. Conf. on Wavelet Analysis and Pattern Recognition*, pp. 373–377 (August 2008)
- [6] Li, J., Lu, B.-L.: A Framework for Multi-view Gender Classification. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) *ICONIP 2007, Part I*. LNCS, vol. 4984, pp. 973–982. Springer, Heidelberg (2008)
- [7] Leng, X.M., Wang, Y.D.: Gender Classification Based on Fuzzy SVM. In: *Int. Conf. on Machine Learning and Cybernetics* (2008)
- [8] Baluja, S., et al.: Boosting Sex Identification Performance. *Int. J. of Computer Vision* 71(1), 111–119 (2007)
- [9] Shen, B.-C., Chen, C.S., Hsu, H.H.: Fast Gender Recognition by Using a Shared Integral Image Approach. In: *IEEE ICASSP* (2009)
- [10] Wang, Z.-H., Mu, Z.-C.: Gender Classification Using Selected Independent-Features Based on Genetic Algorithm. In: *8th Int. Conf. on Machine Learning and Cybernetics*, pp. 394–398 (July 2009)
- [11] Andreu, Y., Mollineda, R.A.: On the Complementarity of Face Parts for Gender Recognition. In: Ruiz-Shulcloper, J., Kropatsch, W.G. (eds.) *CIARP 2008*. LNCS, vol. 5197, pp. 252–260. Springer, Heidelberg (2008)
- [12] Toews, M., Arbel, T.: Detection, Localization, and Sex Classification of Faces from Arbitrary Viewpoints and under Occlusion. *IEEE Trans. on PAMI* 31(9) (September 2009)
- [13] Aghajanian, J., et al.: Within-Object Classification. In: *IEEE 12th ICCV* (2009)
- [14] Li, Z., Zhou, X., Huang, T.S.: Spatial Gaussian Mixture Model for Gender Recognition. In: *IEEE ICIP* (November 2009)
- [15] Cootes, T.F., Wheeler, G.V.: View-Based Active Appearance Models. *Image and Vision Computing* 20, 657–664 (2002)

Multiple Objects Tracking across Multiple Non-overlapped Views

Ke-Yin Chen¹, Chung-Lin Huang¹, Shih-Chung Hsu¹, and I-Cheng Chang²

¹ Department of Electrical Engineering, National Tsing Hua University,
Hsin-Chu, Taiwan
g9761587@oz.nthu.edu.tw, clhuang@ee.nthu.edu.tw,
chvjohfff@gmail.com

² Department of Information Science and Engineering, National Don-Hwa University,
Ha-Lien, Taiwan
ICChang@mail.ndhu.edu.tw

Abstract. This paper introduces a tracking algorithm to track the multiple objects across multiple non-overlapped views. First, we track every single object in each single view and record its activity as the object-based video fragments (OVFs). By linking the related OVFs across different cameras, we may connect two OVFs across two non-overlapped views. Because of scene illumination change, blind region lingering, and objects similar appearance, we may have the problem of path misconnection and fragmentation. This paper develops the Error Path Detection Function (EPDF) and uses the augmented feature (AF) to solve those two problems.

Keywords: Object tracking, Object-based Video Fragment (OVF), Augmented feature (AF), Error Path Detection Function (EPDF).

1 Introduction

Video surveillance system is constructed by a network of cameras with multiple non-overlapped views. In each camera, a period of video of each object's activity is recorded in a so-called object-based video fragment (*OVF*). This paper introduces a method to connect two *OVFs* of the same object moving across two non-overlapped views. Because of scene illumination change, blind region lingering, and objects similar appearance, the system faces the problems of *OVF* misconnection and fragmentation. Our method can detect and correct the miss-connected *OVFs*, and then reconnect the *OVFs* of the same object moving across cameras.

Lee *et al.* [2] proposed an approach for tracking objects in the cameras with overlapped field of views (FOVs) without calibration. Khan *et al.* [3] used *FOV* line constraints for tracking objects in overlapped cameras. Multi-camera tracking approaches with overlapped *FOVs* have been proposed [4, 5]. In non-overlapped views, Kettner *et al.* [6] presented a Bayesian solution to track objects across multiple cameras with non-overlapped views. Porikli *et al.* [7] combined spatiotemporal and appearance cues to track objects and solve the inter-camera color calibration problem.

Black *et al.* [1] used the HSI color space to improve illumination invariance. Javed *et al.* [8] present a camera network topology learning method using the path probabilities of objects. Individual tracks are found by searching the maximal posterior probability of the spatiotemporal and color appearance. Javed *et al.* [9, 10] developed the subspace of inter-camera brightness transfer functions to solve the problem of appearance change across the scenes. D’Orazio *et al.* [11] compared different methods to evaluate the color Brightness Transfer Function (*BTF*) between non overlapped cameras.

Chen *et al.* [12] proposed an unsupervised method to learn both spatiotemporal and appearance relationships for long-term monitoring. They consider the environment changes, such as sudden lighting change. Dick *et al.* [13] used a stochastic transition matrix to describe the observed pattern of people motion within and between FOVs. Ellis *et al.* [14] developed an automatic labeling method to construct the network topology. Stauffer *et al.* [15] built a correspondence model for cameras with both overlapped and non-overlapped FOVs.

Mehmood *et al.* [17] combined the optical flow, feature matching and shape descriptors to detect and track objects efficiently. Their method can be applied to multiple non-overlapped cameras to attain correct inter-camera correspondences. Piccardi *et al.* [18] used the Major Color Spectrum Histogram representation (MCSHR) to represent a moving object. Based on k-means clustering, the reduced color space is used to tolerate the minor changes in color between different cameras and lighting. Song *et al.* [19] combined short term feature correspondence with long-term feature dependency models to derive a path smoothness function for error correspondence correction.

This paper presents a multiple objects tracking across multiple non-overlapped views by using spatio-temporal cues and appearance cues in different views. Our system consists of (1) applying the foreground extraction method to segment the foreground object, (2) using the spatiotemporal cues and appearance connect the related OVFs across different views, (3) using Augmented Feature (*AF*) propagation method to solve the fragmentation and miss-connection problems. Different from [19], our major contributions are proposing the *Error Path Detection Function (EPDF)* to find the miss-connection, and using the *AF* to re-connect the OVFs.

2 Problem Formulation

Our problem is formulated as multiple-object tracking in non-overlapped multiple views. The camera network can be described by a graph of which each node represents the scene of a certain view. As shown in Figure 1, there are six non-overlapped views. Each scene (node) may have more than one *zone*, and each zone can be either an entrance or an exit of the scene. In Figure 1, we find four zones in node 2, and only one zone in nodes 1 and 3. Every two zones are either direct or non-direct related. Two zones in the same node or two neighboring nodes are direct-related, otherwise they are non-direct related. If two direct-related zones are in the same node, they have *intra-zone* relationship, otherwise, they have *inter-zone* relationship.

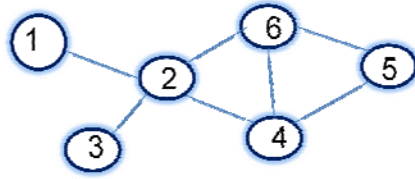


Fig. 1. The camera network topology

Any object moving between two intra-zones in the same view can be tracked and recorded as an object-based video fragment (*OVF*). The object movements between two inter-zones in different views are unknown but predictable. Our goal is to link the related *OVFs* across *inter-zones* by finding the objects in two neighboring views with the similar spatiotemporal cue and appearance cue. The linkage between two *OVFs* is called a *joint*. Here, we assume that (1) the system can track any object within one single view, (2) the cameras are synchronous, (3) each *OVF* is marked by time stamp, object appearance and location information, and (4) the zones in each scene are known priori.

The region (or linkage) between two inter-zones is a *closed blind region*. Figure 1 shows a closed blind region between nodes 1 and 2. Object leaving the exit zone of node 1 will enter the close blind region, and re-appear in the enter zone of node 2 sooner or later. We also define another blind region, called “*open blind region*” in which the object may not necessarily re-appear in any other node. For certain node adjacent to the open blind region, it has no inter-zone relationship with any other node. Object may enter or leave the scenes through the open blind region. Figure 1 also shows an open blind region in node1 or 3.

2.1 Object Tracking in Single View

First, we apply background subtracting and shadow removal to extract the foreground object when it enters the enter zone. Based on the extracted foreground object, the object model can be obtained which can be used for object tracking. In the non-overlapped scenes, each moving object appears in only one single view at any time instance. Here, we apply HS (Hue-Saturation) color histogram to model the object, and then use Mean-Shift algorithm [20] to track the moving object which is enclosed by a rectangle as shown in Figure 2. The rectangle is represented as $s = \{x, y, h, r\}$, where (x, y) represents the center of the rectangle, and (h, r) represents the height and the aspect ratio.



Fig. 2. The two video fragments of the same object

2.2 Spatiotemporal and Appearance Cues

The appearance cue of each object is modeled by the HS (Hue-Saturation) color histogram of the rectangle enclosing the moving object. The similarity measure between the observations of two objects is described by computing Bhattacharyya coefficient ρ based on the color histograms, $p(u)=\{p^{(u)}\}_{u=1\dots m}$ and $q(u)=\{q^{(u)}\}_{u=1\dots m}$ of two objects. Larger ρ indicates more similar between these two color histograms. The similarity distance between two objects is measured by $d = \sqrt{1 - \rho[p(u), q(u)]}$. The color distribution of each object is temporally updated.

By exploiting the camera network topology, we can describe the spatiotemporal relationship between the cameras in terms of the *transition time* and the *transition probability*. The former indicates the time duration for an object moving from one exit zone to the other entry zone, and latter is the probability distribution of the transition time between two observations in two inter-zones. The spatiotemporal relationship between two inter-zones is based on the camera network topology. After the training phase, we have the transition probability for each possible linkage between two inter-zones. For inter-zones a and b of two different views, we use $P_{ab}(T)$ to describe the transition probability that people move from zone a to zone b after time T . The same object exits from zone a at time T_i and enters zone b at time T_j , then $T=T_j-T_i$.

3 Inter-zone Video Fragments Linkage

Here, we use the spatiotemporal and appearance cues of the observations to generate a preliminary linkage of *OVFs* across inter-zone. Based on the observations of *OVFs* across inter-zones, we may create the linkage of the two *OVFs*. For each zone, there is a handover list. The handover list of zone a (i.e., H_a) is defined as the collection of the observations of the objects appearing in the adjacent zones of zone a .

Object A enters the zone Z_A with the observation denoted as O_A which consists of the spatiotemporal cue $O_A(st)$ and the appearance cue $O_A(app)$. The $O_A(st)$ includes the camera id , the zone id , the position, and time of appearance at the zone A as $T(O_{Z_A})$. Then we find the best corresponding object with the observation O_h in the handover list H_A . Based on $O_A(st)$ and $O_A(app)$, we find the best matched one in H_A . If the highest probability exceeds a threshold, we label the new observation O_A and the observation O_h as the same object. Otherwise, object A is treated as a new object entering in the scene. The similarity between the observation O_A and the related one O_h in the handover list H_A (i.e., $O_h \in H_A$) is described as $p(O_A, O_h)$. The most likely one in H_a can be obtained as

$$\varphi = Arg \max_h p(O_A, O_h) \quad (1)$$

Assuming $O_A(st)$ and $O_A(app)$ are independent so that we can compute likelihood of similarity based on $O_A(st)$ and $O_A(app)$ with different weights. Equation (1) can be rewritten as

$$\begin{aligned} \varphi &= \text{Arg max}_h p(O_A, O_h) \\ &= \text{Arg max}_h [w \cdot p(O_A(st), O_h(st)) + (1-w) p(O_A(app), O_h(app))] \end{aligned}$$

where $p(O_A(app), O_h(app))$ is the probability of appearance similarity, and $p(O_A(st), O_h(st))$ is the probability of spatiotemporal similarity defined as

$$p(O_A(st), O_h(st)) = \sum_{\forall Z_A} \sum_{\forall Z_h} P_{Z_A Z_h}(T) [p(O_A(st)|Z_A) p(O_h(st)|Z_h)] \tag{2}$$

where $P_{Z_A Z_h}(T)$ is the transition probability of travel time between two inter-zones Z_A and Z_h as $T = T[\mathbf{O}_{Z_A}] - T[\mathbf{O}_{Z_h}]$, and $P(\mathbf{O}^* | Z^*)$ is the probability of the observation \mathbf{O}^* entering or exiting from zone Z^* . To connect two *OVFs*, we dynamically adjust the weighting for spatiotemporal features and appearance features. If the illumination changes make the appearance features unreliable for similarity measure, we will increase the weight of spatiotemporal features.

Connecting *OVFs* can be viewed as a labeling problem. Two *OVFs* assigned with the same label will be linked together indicating the activity of the same individual object. The initially cascaded *OVFs* are called the *OVF link*. In Figure 3, we show the ground truth of two *OVF links* for two moving objects. However, we may have six initial *OVF links*. The path fragmentation problem occurs when the connection between *OVFs* fails because of (1) two or more people with the same appearance, and (2) the lingering time of the designated object is longer than the others.

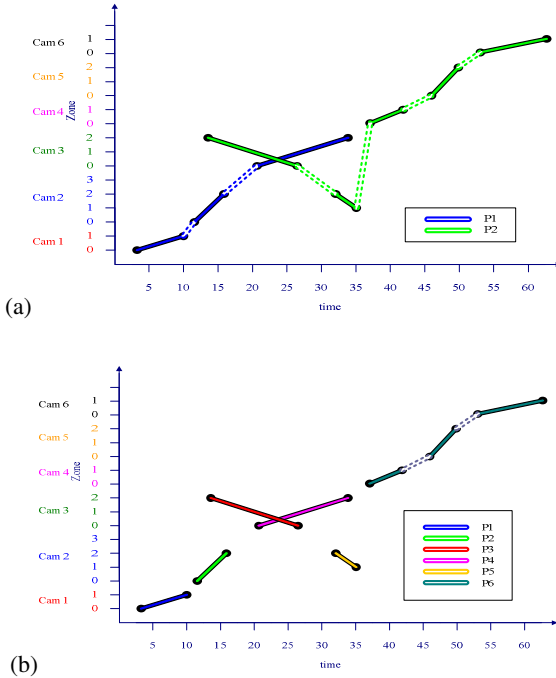


Fig. 3. (a) Ground truth, (b) six linked paths

The initially *OVF links* may have two problems: *path fragmentation* and *path misconnection*. The ***path fragmentation*** indicates that the *OVFs* of the same object are not connected across the inter-zones. It occurs due to *variant lighting* and *uncertain lingering time* in blind region. The ***path misconnection*** indicates that the *OVFs* of different objects may be miss-connected together. It occurs when people dressed in similar clothes may appear from the blind region at the same time, and appearance cue becomes not reliable. Therefore, we need to find and correct the miss-connection and solve the fragmentation.

4 Error Linkage Correction

Because of the different viewing angles and positions of the cameras, the observations from different cameras are not the same. In addition to the spatiotemporal and appearance cues, we may have another feature, the human face. The human face feature can be detected and treated as the augmented feature (*AF*) for correcting the path miss-connection. The correction process consists of four steps: (1) Calculate the error path detection function (*EPDF*) at the joints to check the validity of the linked *OVFs*, (2) Divide the *OVF link* is divided into two *OVF sub-links* at the joint if there is an error, (3) Propagate the *AF* in the same *OVF sub-link*, and (4) Re-calculate the similarity between the *OVF sub-links* for path correction.

4.1 Misconnection Detection

Path misconnection usually occurs when several objects with similar appearance pass through closed blind region at the same time. Here we propose the Error Path Detection Function (*EPDF*) to represent the possible misconnection. Two *OVFs* has been connected at joint and assigned to the same link L_z as the $i-1^{th}$ and i^{th} fragment as $O_{z,i-1}$ and $O_{z,i}$. We compare the possible connection of $O_{z,i-1}$ and O_b with the proposed one, in which Q_b in the handover list of $Q_{z,i-1}$ as $Q_b \in H(Q_{z,i-1})$. If the difference is not large enough, then the connection may not be correct. We use *EPDF* to identify the reliability of the connection (or *joint*) as

$$EPDF(L_{z,i}) = \begin{cases} 0 & \text{if } |P(O_{z,i-1}, O_{z,i}) - P(O_{z,i-1}, O_b)| > Thres \\ 1 & \text{Otherwise} \end{cases}$$

where $L_{z,i}$ is defined as the i th joint of link z , $i=1, \dots, N_z-1$, and N_z represents the number of fragments in the link.

Figure 4 shows an example of three objects of which Object 1 and Object 3 have similar color appearance. Object 3 leaves zone 0 of camera 3 first and then Object 1 leaves later. We compute *EPDF* at the joint of every *OVF link* to find the error connection. Figure 5 shows the *EPDF* of three *OVF links* respectively, of which two *OVF links* have misconnection problem. The 1st *OVF link* indicates that there are another similar appearance objects in the blind region simultaneously, so that the difference value is less than threshold. We calculate *EPDF* of every *OVF link* to find the misconnected joint.

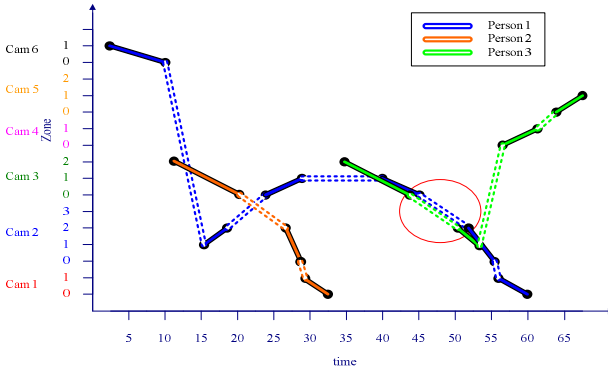


Fig. 4. The initial *OVF* links with misconnection

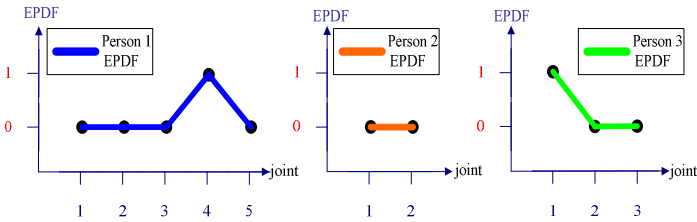


Fig. 5. The EPDF of three links

4.2 Augmented Feature Propagation

The captured observation consists of spatiotemporal and appearance cues. For some observations which cannot be obtained in each view are called the Augmented Features (*AF*). Since the camera viewing directions are different, the *AF* may not be found in every connected *OVF*. Since the connected *OVFs* are supposed to have the same *AF*, we may propagate *AF* across the connected *OVFs*.

We propagate the *AFs* to all *OVFs* in the same *OVF* link. The *AFs* of *OVF* links are used to calculate their similarity. We compare every two *OVF* sub-links, and then connect the two *OVF* sub-links with the highest similarity. The path misconnection and fragmentation problem can be solved by the following steps:

- (1) For each *OVF* link L_z calculate *EPDF* for each joint i .
- (2) Segment link L_z into a *OVF* sub-links S_x and S_y .
- (3) Propagate the additional *AFs* to the other fragments of the same *OVF* sub-link.
- (4) Establish the correspondence between the observations of every two *OVF* sub-links.
- (5) If there is only one *OVF* sub-link in handover list, they can be connected directly.
- (6) For each cascaded *OVF* link, re-compute the *EPDF* at every joint.
- (7) Repeat the above steps until *EPDF* of this path is zero, or else it fails.

The correspondence between two *OVF* sub-links S_{ax} and S_{by} can be obtained based on the observations Q_{ax} and Q_{by} . The likelihood of the two observations is described as $p(Q_{ax}, Q_{by})$ with S_{by} in the handover list of S_{ax} as $S_{by} \in H(S_{ax})$. Assume that the spatiotemporal cue, the appearance cue and the augmented cue are independent. The most likely corresponding *OVF* sub-links can be described as

$$\varphi = \text{Argmax}_{by} p(Q_{a,x}(aug), Q_{b,y}(aug))$$

where $p(Q_{a,x}(aug), Q_{b,y}(aug))$ is likelihood of the two observations based on the *AF* similarity.

Figure 6 shows the results of *AF* Propagation. The *OVF* link L_1 is divided into two *OVF* sub-links S_{11} and S_{12} , and *OVF* link L_2 is also divided into two *OVF* sub-links S_{21} and S_{22} . The *AF* is propagated in every *OVF* sub-link. In *OVF* sub-link S_{22} , the *AF* of *OVF* #2 is propagated from *OVF* #4. Therefore, S_{22} , S_{11} , and S_{21} have the same similar *AF*s. Based on the propagated *AF*s, we may compute the similarity between two *OVF* sub-links. Each *OVF* sub-link will be connected to another *OVF* sub-link with the largest similarity.

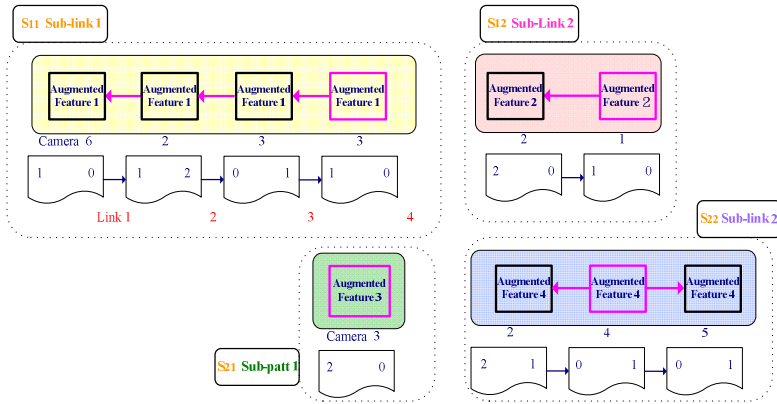


Fig. 6. The Propagation of AFs in *OVF* sub-links

Since the similarity between S_{22} and S_{11} is much larger, S_{22} is cascaded with S_{11} , as a new link which will retain *AF* #1 and *AF* #4 simultaneously. Once the connection is determined, the *EPDF* of 4th joint of the new link of will become zero.

There is only one *OVF* sub-link S_{21} in handover list, so that it is connected with S_{12} and become a new *OVF* link. The *EPDF* of 1st joint will be zero. Due to similar color appearance of different objects, the wrong linkages and cascaded *OVF* links are generated. As shown in Figure 7, *OVF* link 2 has two incorrect linkages, which are indicated by a pink circle.

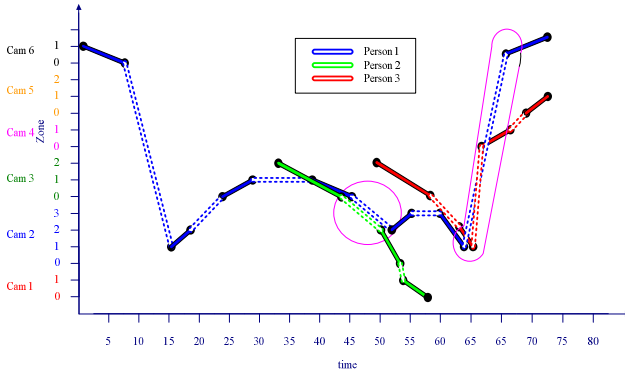


Fig. 7. Initial miss-connected linkages

Figure 8 shows that an *AF* (e.g., human faces features) is propagated to other *OVFs* in the same sub-link. Solid lines represent the initial links, and the dashed lines indicate the ground truth. Sub-link S_{22} is miss-connected with S_{21} . Sub-links S_{32} and S_{12} are not linked because of no *AF* propagation. S_{22} is more coherent with S_{11} than with S_{21} , $p(Q_{22}, Q_{11}) > p(Q_{22}, Q_{21})$. Therefore, S_{22} is connected with S_{11} , and the *EPDF* is set as zero. There is only one sub-link S_{21} in S_{12} handover list, so they are connected.

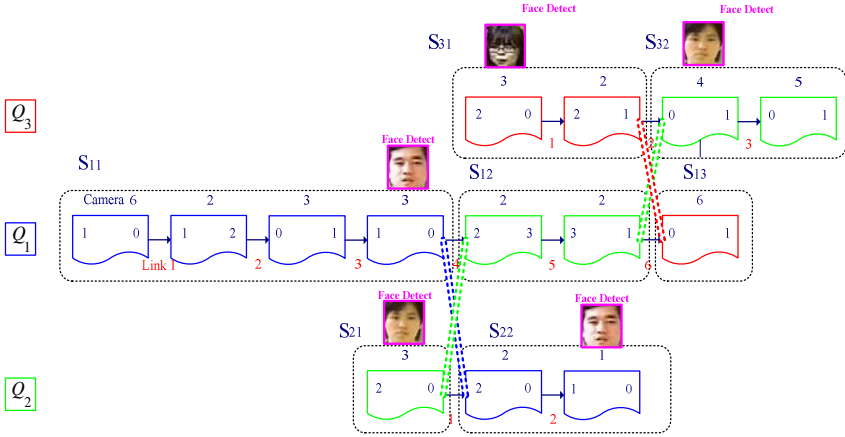


Fig. 8. The reconnection between two *OVF* sub-links

5 Experimental Results

In the experiments, we have the synchronized videos from six non-overlapped cameras. The format of image frame is $320 \times 240 \times 24$ bits and the frame rate is 25 frames/sec. Figure 9 shows six indoor non-overlapped views in the experiment. The color histogram of the object is used as the basic feature for object tracking. Each tracked object in each view is enclosed by a rectangle block. The parameters of each block are the position, the height, and the aspect ratio of the block.

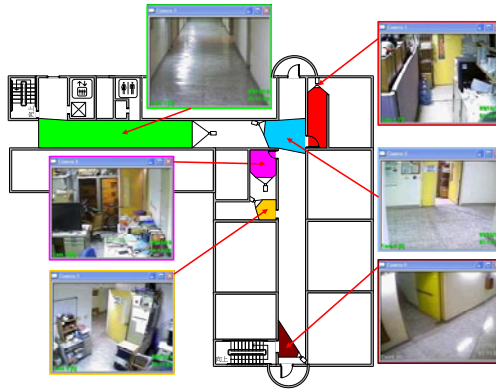


Fig. 9. The non-overlapped multi-cameras system

To illustrate the effectiveness of our system, we demonstrate three experiments.

a) Experiment 1. Three people enter in the viewing of camera 6 individually. They walk together in the blind region at the same time, and then leave the blind region individually. Each object can be tracked independently after the walking through the blind region.

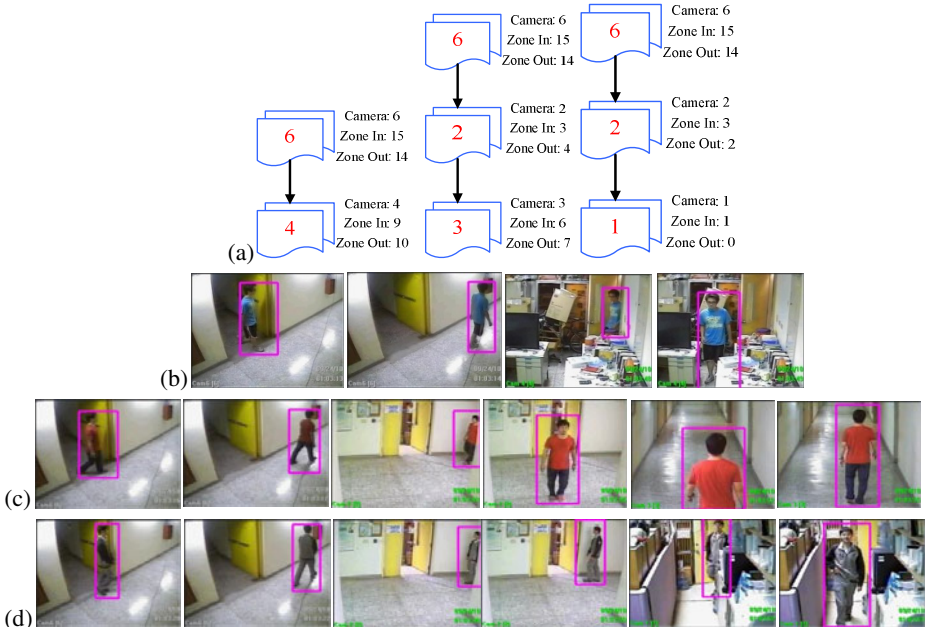


Fig. 10. The experimental results of experiment #1. (a) three different paths, (b)~(d) the frames of each OVf links.

b) Experiment 2. Two people enter the scene of camera3 individually. Object 2 leaves camera 3 first, and object 1 leaves later. But object 1 enters the scene of camera 2 first, object 3 enters later. Their spatiotemporal similarity is close, and they have similar appearances. In the initial *OVF* linkage, path miss-connection occurs when two objects leave the blind region. The *OVF* link of object 2 in camera 2 and 1 will be connected to the *OVF* link of object 1 in camera 3. The *OVF* links of object 1 in cameras 2, 4 and 5 will be connected to the *OVF* link of object 2 in camera 3.

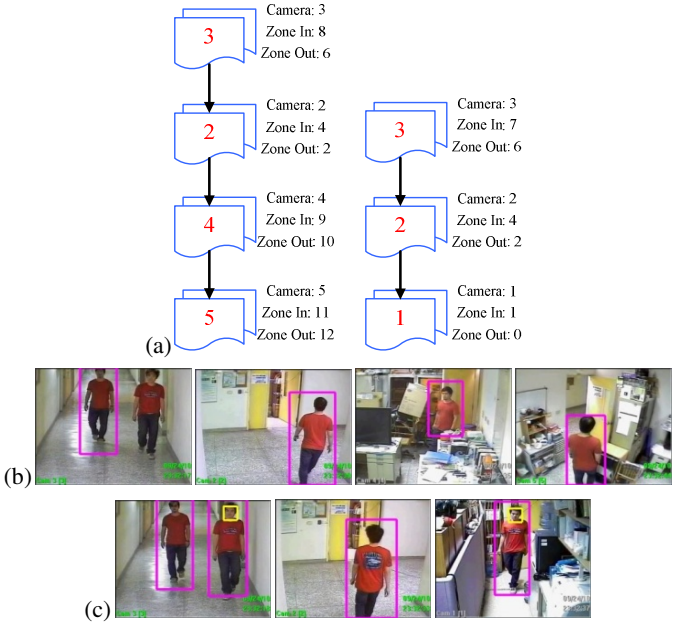


Fig. 11. The Experimental results of Experiment #2. (a) two different paths, (b)~(c) the frames of each *OVF* links.

c) Experiment 3. Three people with similar appearances (dressed in the same colors) appear in the scene. When they enter the same closed blind region, the path miss-connection problem occurs. *OVF* link #2 is a miss-connection. We employ the *EPDF* to find the miss-connected joints and divide the miss-connected *OVF* links into two *OVF* sub-links. By applying *AF* propagation, we can reconnect the *OVF* sub-links as *OVF* link.

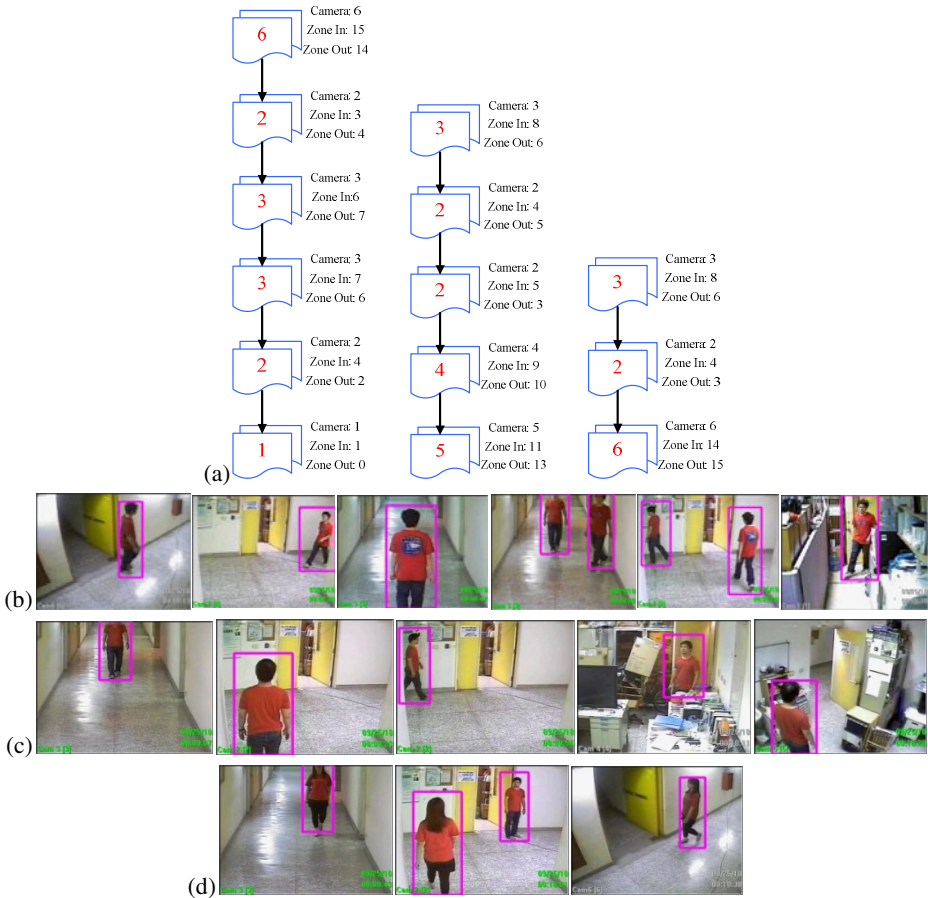


Fig. 12. The Experimental results of Experiment #3. (a) three different paths, (b)~(d) the frames of each OVF links.

6 Conclusions

The paper presents a tracking system for multiple cameras with non-overlapped views by exploiting the basic features, spatiotemporal features, and appearance features to determine human's tracks across cameras. We have shown that our method can detect and correct the miss-connected *OVFs*, and then reconnect the *OVFs* of the same object moving across cameras.

References

- [1] Black, J., et al.: Wide Area Surveillance with a Multi-Camera Network. In: Proc. of Intelligent Distributed Surveillance Systems (2003)
- [2] Lee, L., et al.: Monitoring Activities from Multiple Video Streams: Establishing a Common Coordinate Frame. IEEE Trans. PAMI 22(8), 758–768 (2000)

- [3] Khan, S., et al.: Consistent Labeling of Tracked Objects in Multiple Cameras with Overlapped Fields of View. *IEEE Trans. PAMI* (2003)
- [4] Javed, O., et al.: KNIGHTM: a real time surveillance system for multiple overlapped and non-overlapped cameras. In: *ICME* (2003)
- [5] Zhu, L.-J., et al.: Tracking of multiple objects across multiple cameras with overlapped and non-overlapped views. In: *IEEE ISCAS* (2009)
- [6] Kettner, V., et al.: Bayesian Multi-camera Surveillance. In: *IEEE CVPR* (1999)
- [7] Porikli, F., et al.: Multi-Camera Calibration, Object Tracking and Query Generation. In: *IEEE ICME* (2003)
- [8] Javed, O., et al.: Tracking across Multiple Cameras with Disjoint Views. In: *9th IEEE ICCV* (October 2003)
- [9] Javed, O., et al.: Appearance modeling for tracking in multiple non-overlapped cameras. In: *IEEE CVPR 2005*, vol. 2, pp. 26–33 (June 2005)
- [10] Javed, O., et al.: Modeling inter-camera space–time and appearance relationships for tracking across non-overlapped views. *Computer Vision and Image Understanding*, 146–162 (2008)
- [11] D’Orazio, T., et al.: Color Brightness Transfer Function Evaluation for Non overlapped Multi Camera Tracking. In: *ICDSC* (2009)
- [12] Chen, K.W., et al.: An Adaptive Learning Method for Target Tracking across Multiple Cameras. In: *IEEE CVPR 2008*, pp. 1–8 (June 2008)
- [13] Dick, A., et al.: A Stochastic Approach to Tracking Objects across Multiple Cameras. In: *Australian Conf. on Artificial Intelligence*, pp.160–170 (2004)
- [14] Ellis, T.J., et al.: Learning a Multi-Camera Topology. In: *IEEE Workshop on VS-PETS* (2003)
- [15] Stauffer, C.: Learning to Track Objects through Unobserved Regions. In: *IEEE Workshop on Motion and Video Computing*, pp. 96–102 (January 2005)
- [16] Mehmood, M.O.: Multi-camera based Human Tracking with Non- Overlapped Fields of View. In: *Int. Conf. on AICT 2009*, pp. 1–6 (October 2009)
- [17] Cheng, E.D., et al.: Mitigating the Effects of Variable Illumination for Tracking across Disjoint Camera Views. In: *IEEE AVSS 2006* (November 2006)
- [18] Piccardi, M., et al.: Track matching over disjoint camera views based on an incremental major color spectrum histogram. In: *IEEE AVSS* (2005)
- [19] Song, B., et al.: Robust Tracking in A Camera Network A Multi-Objective Optimization Framework. *IEEE J. on Selected Topics in Signal Processing*, 582–596 (2008)
- [20] Comaniciu, D., Meer, P.: Mean-Shift: A robust Approach toward feature space analysis. *IEEE Trans. on PAMI* 24(5), 603–619 (2002)

Fast Hypercomplex Polar Fourier Analysis for Image Processing

Zhuo Yang and Sei-ichiro Kamata

Graduate School of Information, Production and Systems

Waseda University, Japan

joel@ruri.waseda.jp, kam@waseda.jp

Abstract. Hypercomplex polar Fourier analysis treats a signal as a vector field and generalizes the conventional polar Fourier analysis. It can handle signals represented by hypercomplex numbers such as color images. It is reversible that can reconstruct image. Its coefficient has rotation invariance property that can be used for feature extraction. With these properties, it can be used for image processing applications like image representation and image understanding. However in order to increase the computation speed, fast algorithm is needed especially for image processing applications like realtime systems and limited resource platforms. This paper presents fast hypercomplex polar Fourier analysis that based on symmetric properties and mathematical properties of trigonometric functions. Proposed fast hypercomplex polar Fourier analysis computes symmetric eight points simultaneously that significantly reduce the computation time.

Keywords: fast hypercomplex polar Fourier analysis, hypercomplex polar Fourier analysis, Fourier analysis.

1 Introduction

Fourier transforms have been widely used in image processing, signal processing and many engineering fields [1]. By representing image as hypercomplex numbers, especially the quaternions discovered by Hamilton [2], HyperComplex Fourier transform is proposed as generalization of quaternion Fourier transform for color image processing [3]. The relationship between right-side quaternion Fourier transform and left-side quaternion Fourier transform is established [4]. Based on hypercomplex Fourier transform, effective algorithms for motion estimation in color image sequences are studied [5]. Quaternionic Gabor filters are designed to combine the color channels and the orientations in the image plane [6].

Inspired from these, hypercomplex polar Fourier analysis is studied. By introducing a hypercomplex number, hypercomplex polar Fourier analysis [8] treats a signal as a vector field and generalizes the polar Fourier analysis [7]. Hypercomplex polar Fourier analysis can handle color image. With orthogonality, it can decompose and reconstruct color image. The coefficients hold rotation invariant property. With these properties, it can be widely used as an image

processing tool. Unfortunately, the computations of hypercomplex polar Fourier analysis involve many Bessel function and trigonometric computations that no fast method has been reported. Therefore, reduction of the computation time is very significant.

This paper focuses on fast hypercomplex polar Fourier analysis. Fast and compact method to compute the coefficients of hypercomplex polar Fourier analysis is proposed by using mathematical properties of trigonometric functions and points relationships. The two dimensional basis function of hypercomplex polar Fourier analysis has symmetry properties with respect to the x axis, y axis, $y = x$ line, $y = -x$ line and origin that can be used for fast computation. The computational complexity can be reduced by calculating half of the first quadrant. For image processing applications, computation time is important factor. Using the proposed method, only one eighth is needed compared with the direct calculation.

The organization of this paper is as follows. The basic theory of hypercomplex polar Fourier analysis including mathematics definitions are provided in Section 2. Section 3 presents the proposed method in detail. Experiments are designed to demonstrate effectiveness of the proposed method in Section 4. Finally, Section 5 concludes this study.

2 Background

2.1 Hypercomplex Number

As a type of hypercomplex number and generalization of complex number, the quaternion, its properties and applications have been studied [9]. In signal and image processing, quaternion number based methods are actively researched. Such as, quaternionic Gabor filters are designed to combine the color channels and the orientations in the image plane [6]. Quaternionic phase correlation based motion estimation approach is studied [5].

Complex number has two components, the real part and imaginary part. Quaternion has one real part and three imaginary parts. Given $a, b, c, d \in \mathbb{R}$, a quaternion $q \in \mathbb{H}$ (\mathbb{H} denotes Hamilton) is defined as

$$q = \mathcal{S}(q) + \mathcal{V}(q), \mathcal{S}(q) = a, \mathcal{V}(q) = bi + cj + dk \quad (1)$$

where $\mathcal{S}(q)$ is scalar part and $\mathcal{V}(q)$ is vector part. i, j, k are imaginary operators obeying the following rules

$$\begin{aligned} i^2 = j^2 = k^2 &= -1, ij = -ji = k, \\ jk &= -kj = i, ki = -ik = j, \end{aligned} \quad (2)$$

The norm of quaternion q is

$$\|q\| = \sqrt{a^2 + b^2 + c^2 + d^2}. \quad (3)$$

Quaternion q is named as unit quaternion if it is in set

$$\mathbb{U} = \{q | q \in \mathbb{H}, \|q\| = 1\}. \quad (4)$$

If quaternion q in following set,

$$\mathbb{P} = \{q|q \in \mathbb{H}, \mathcal{S}(q) = 0\}, \tag{5}$$

it is called pure quaternion. The quaternions belonging to set

$$\mathbb{S} = \{q|q \in \mathbb{U}, q \in \mathbb{P}\}, \tag{6}$$

are called unit pure quaternion. Euler formula holds for hypercomplex numbers,

$$e^{\mu\phi} = \cos(\phi) + \mu \sin(\phi) \tag{7}$$

2.2 HyperComplex Polar Fourier Analysis

Given a 2D function $f(x, y)$, it can be transformed from cartesian coordinate to polar coordinate $f(r, \varphi)$, where r and φ denote radius and azimuth respectively. The following equations transform from cartesian coordinate to polar coordinate,

$$r = \sqrt{x^2 + y^2}, \tag{8}$$

and

$$\varphi = \arctan \frac{y}{x}. \tag{9}$$

Hypercomplex Polar Fourier analysis involves points within the largest inner circle of the image. After normalization, it is defined on the unit circle that $r \leq 1$ and can be expanded with respect to the basis function. Hypercomplex polar Fourier analysis is defined as

$$f(r, \varphi) = \sum_{n=1}^{\infty} \sum_{m=-\infty}^{\infty} HP_{nm} R_{nm}(r) e^{\mu m \varphi}, \tag{10}$$

where μ is unit pure quaternion and is defined as $\mu = \frac{1}{\sqrt{3}}i + \frac{1}{\sqrt{3}}j + \frac{1}{\sqrt{3}}k$, and the coefficient is

$$\begin{aligned} HP_{nm} &= \frac{1}{\sqrt{2\pi}} \int_0^1 \int_0^{2\pi} R_{nm}(r) f(r, \varphi) e^{-\mu m \varphi} r dr d\varphi \\ &= \frac{1}{\sqrt{2\pi}} \int_0^1 \int_0^{2\pi} R_{nm}(r) f(r, \varphi) (\cos m\varphi - \mu \sin m\varphi) r dr d\varphi \end{aligned} \tag{11}$$

where

$$R_{nm}(r) = \frac{1}{\sqrt{N_n^{(m)}}} J_m(x_{mn}r), \tag{12}$$

in which J_m is the m -th order first class Bessel series [10], and $N_n^{(m)}$ can be deduced by imposing boundary conditions according to the Sturm-Liouville(S-L) theory [11]. With zero-value boundary condition,

$$N_n^{(m)} = \frac{1}{2} J_{m+1}^2(x_{mn}), \tag{13}$$

in which x_{mn} is the n th positive root for $J_m(x)$.

The coefficient HP_{nm} is rotation invariant. Hypercomplex polar Fourier analysis is reversible. Fig. 1 shown that image can be reconstructed. With n increases bigger, more detail part of the image can be obtained.

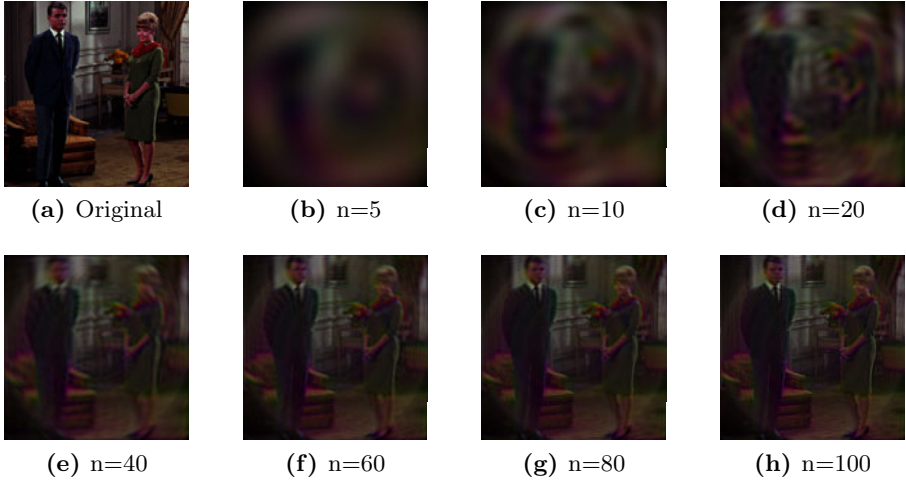


Fig. 1. Hypercomplex polar Fourier analysis

3 Fast Hypercomplex Polar Fourier Analysis

This section presents fast hypercomplex polar Fourier analysis. From Eq. 11, for same radius r , the different integrand part for each point is $f(r, \varphi)(\cos m\varphi - \mu \sin m\varphi)$. As shown in Fig.2 , point (x, y) is a point in first quadrant below $y = x$, has seven other symmetric points with respect to x axis, y axis, $y = x$, $y = -x$ and origin.

Mappings between polar and cartesian coordinates are show in Table 1.

Within period 2π , $\sin(\varphi)$ and $\cos(\varphi)$ functions are periodic functions. Periods for $\sin(m\varphi)$ and $\cos(m\varphi)$ are $2\pi/m$. Derived from the periodic and symmetric properties of trigonometric functions that used in FFT [12] , mathematical relationships for trigonometric functions exist with respect to different m . If l is divided by 4 with remainder 3 that means $\text{mod}(l, 4) = 3$, following relationship for sine function can be deduced

$$\sin(l(\frac{\pi}{2} - \theta)) = -\cos(l\theta), \tag{14}$$

$$\sin(l(\frac{\pi}{2} + \theta)) = -\cos(l\theta), \tag{15}$$

$$\sin(l(\pi - \theta)) = \sin(l\theta), \tag{16}$$

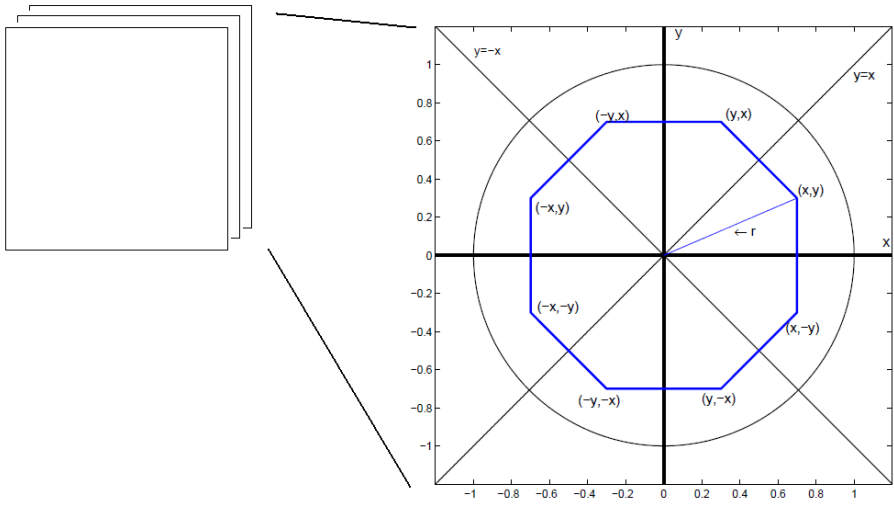


Fig. 2. Two dimensional symmetric points in multiple layers

$$\sin(l(\pi + \theta)) = -\sin(l\theta), \tag{17}$$

$$\sin(l(\frac{3\pi}{2} - \theta)) = \cos(l\theta), \tag{18}$$

$$\sin(l(\frac{3\pi}{2} + \theta)) = \cos(l\theta), \tag{19}$$

$$\sin(l(2\pi - \theta)) = -\sin(l\theta). \tag{20}$$

Similar relationships also exist for cosine function and other l values. For the eight symmetric points on the same radius r , if coefficients can be calculated simultaneously, then the computation time for trigonometric function and Bessel function can be reduced.

Based on foregoing discussion, fast hypercomplex polar Fourier analysis is given by

$$FastHP_{nm} = \frac{1}{\sqrt{2\pi}} \iint_D R_{nm}(\sqrt{x^2 + y^2}) (G_m(x, y) - \mu H_m(x, y)) dx dy, \tag{21}$$

where

$$D = \{(x, y) | 0 \leq x \leq 1, 0 \leq y \leq x, 0 \leq x^2 + y^2 \leq 1\}, \tag{22}$$

and $G_m(x, y)$ and $H_m(x, y)$ are given in followings.

With same result, proposed method can share computation between symmetric points that significantly reduce the time in order to obtain the final result. The proposed method is unrelated to image content. Experiments are designed and results are given in following section.

Table 1. (r, θ) and its symmetric points

Polar Coordinate	Cartesian Coordinate
(r, θ)	(x, y)
$(r, \frac{\pi}{2} - \theta)$	(y, x)
$(r, \frac{\pi}{2} + \theta)$	$(-y, x)$
$(r, \pi - \theta)$	$(-x, y)$
$(r, \pi + \theta)$	$(-x, -y)$
$(r, \frac{3\pi}{2} - \theta)$	$(-y, -x)$
$(r, \frac{3\pi}{2} + \theta)$	$(y, -x)$
$(r, 2\pi - \theta)$	$(x, -y)$

$$G_m(x, y) = \begin{cases} \begin{cases} (f(x, y) + f(y, x) + f(-y, x) + f(-x, y) \\ + f(-x, -y) + f(-y, -x) + f(y, -x) + f(x, -y))\cos(m\varphi) & \text{if } \text{mod}(m, 4) = 0 \\ (f(x, y) - f(-x, y) - f(-x, -y) + f(x, -y))\cos(m\varphi) \\ + (f(y, x) - f(-y, x) - f(-y, -x) + f(y, -x))\sin(m\varphi) & \text{if } \text{mod}(m, 4) = 1 \end{cases} \\ \begin{cases} (f(x, y) - f(y, x) - f(-y, x) + f(-x, y) \\ + f(-x, -y) - f(-y, -x) - f(y, -x) + f(x, -y))\cos(m\varphi) & \text{if } \text{mod}(m, 4) = 2 \\ (f(x, y) - f(-x, y) - f(-x, -y) + f(x, -y))\cos(m\varphi) \\ - (f(y, x) - f(-y, x) - f(-y, -x) + f(y, -x))\sin(m\varphi) & \text{if } \text{mod}(m, 4) = 3 \end{cases} \end{cases} \quad (23)$$

$$H_m(x, y) = \begin{cases} \begin{cases} (f(x, y) - f(y, x) + f(-y, x) - f(-x, y) \\ + f(-x, -y) - f(-y, -x) + f(y, -x) - f(x, -y))\sin(m\varphi) & \text{if } \text{mod}(m, 4) = 0 \\ (f(x, y) + f(-x, y) - f(-x, -y) - f(x, -y))\sin(m\varphi) \\ + (f(y, x) + f(-y, x) - f(-y, -x) - f(y, -x))\cos(m\varphi) & \text{if } \text{mod}(m, 4) = 1 \end{cases} \\ \begin{cases} (f(x, y) + f(y, x) - f(-y, x) - f(-x, y) \\ + f(-x, -y) + f(-y, -x) - f(y, -x) - f(x, -y))\sin(m\varphi) & \text{if } \text{mod}(m, 4) = 2 \\ (f(x, y) + f(-x, y) - f(-x, -y) - f(x, -y))\sin(m\varphi) \\ - (f(y, x) + f(-y, x) - f(-y, -x) - f(y, -x))\cos(m\varphi) & \text{if } \text{mod}(m, 4) = 3 \end{cases} \end{cases} \quad (24)$$

4 Experimental Results

The performance of the proposed fast hypercomplex polar Fourier analysis in computation reduction is validated through comparative experiments using different images. Images with different content are tested for test to illustrate the efficiency and feasibility of the proposed method over direct computation. PC

environment (Celeron 1.86GHz, 2G Memory) is used to perform the experiments. Algorithms are implemented by C++. GNU Scientific Library [13] is used for Bessel function.

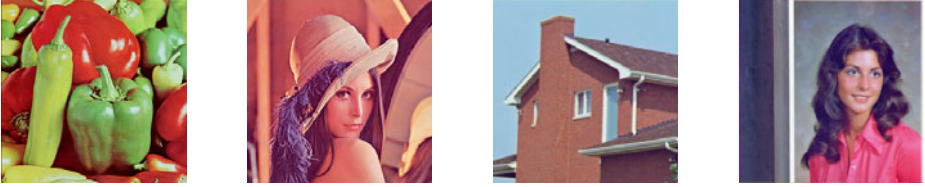


Fig. 3. Standard Images

Test data set consists of standard images as shown in Fig. 3 . With different number of coefficients computed, the performances in terms of CPU elapsed time are given in Table 2 . With same computation result, but two methods have different running time. From the result, the proposed method is effective and is unrelated to number of coefficients and image content. It takes one eighth time to compute hypercomplex polar Fourier analysis that significantly boost the speed. Applications that need hypercomplex polar Fourier analysis for image representation and image understanding can largely reduce the computation time.

Table 2. CPU Elapsed Time to Test Standard Images

Coefficients	Direct Method	Proposed Method	Ratio
5	0.965	0.130	0.135
10	1.924	0.262	0.136
20	3.888	0.532	0.137
30	5.884	0.804	0.137
40	7.846	1.066	0.136

5 Conclusions

In this paper, fast hypercomplex polar Fourier analysis is proposed. Hypercomplex polar Fourier analysis can be used in applications like image representation and image understanding as discussed in previous work. By using the symmetric properties and mathematical properties of trigonometric functions , the proposed methods only calculate one eighth of trigonometric functions and Bessel functions. That means, proposed method can largely decrease the calculation time by sharing computation between symmetric points. Experimental results are given on different images to illustrate the effectiveness. Image processing applications that need fast hypercomplex polar Fourier analysis will benefit from this work.

Acknowledgments. This work was supported in part by Grant-in-Aid (No.21500181) for Scientific Research by the Ministry of Education, Science and Culture of Japan.

References

1. Bracewell, R.N.: The Fourier transform and its applications. McGraw-Hill (2000)
2. Hamilton, W.R.: Elements of Quaternions. Longmans Green, London (1866)
3. Ell, T.A., Sangwine, S.J.: Hypercomplex Fourier Transforms of Color Images. *IEEE Trans. IP* 16(1), 22–35 (2007)
4. Yeh, M.-H.: Relationships Among Various 2-D Quaternion Fourier Transforms. *IEEE Trans. SPL* 15, 669–672 (2008)
5. Alexiadis, D.S., Sergiadis, G.D.: Motion estimation, segmentation and separation, using hypercomplex phase correlation, clustering techniques and graph-based optimization. *Computer Vision and Image Understanding* 113(2), 212–234 (2009)
6. Subakan, O.N., Vemuri, B.C.: A Quaternion Framework for Color Image Smoothing and Segmentation. *International Journal of Computer Vision* 91(3), 233–250 (2011)
7. Wang, Q., Ronneberger, O., Burkhardt, H.: Rotational Invariance Based on Fourier Analysis in Polar and Spherical Coordinates. *IEEE Trans. PAMI* 31(9), 1715–1722 (2009)
8. Yang, Z., Kamata, S.: Hypercomplex Polar Fourier Analysis for Image Representation. *IEICE Trans. on Info. and Sys.* E94-D(8), 1663–1670 (2011)
9. Ward, J.P.: Quaternions and Cayley Numbers, Algebra and Applications. Springer, Heidelberg (1997)
10. Andrews, L.: Special Functions of Mathematics for Engineers, 2nd edn. SPIE Press (1997)
11. Kosmala, W.: Advanced Calculus: a friendly approach. Prentice-Hall (1999)
12. Burrus, C.S., Parks, T.W.: DFT/FFT and Convolution Algorithms and Implementation. John Wiley & Sons (1985)
13. The Gnu Scientific Library, <http://www.gnu.org/software/gsl/>

Colorization by Landmark Pixels Extraction

Weiwei Du, Shiya Mori, and Nobuyuki Nakamori

Information Science, Kyoto Institute of Technology,
Kyoto, Japan 606-8585
{duwewei,nakamori}@kit.ac.jp,
m0651024@edu.kit.ac.jp

Abstract. A one-dimensional luminance scalar is replaced by a vector of a colorful multi-dimension for every pixel of a monochrome image, it is called as colorization. Obviously, it is under-constrained. Some prior knowledge is considered to be given to the monochrome image. Colorization using optimization algorithm is an effective algorithm for the above problem. Scribbles are considered as the prior knowledge. However, it cannot effectively do with complex images without repeating experiments for confirming the place of scribbles. Therefore, in our paper, landmark pixels are considered as the prior knowledge. We propose an algorithm which is colorization by landmark pixels extraction. It need not repeat experiments and automatically generates landmark pixels like scribbles. Finally, colorize the monochrome image according to requirements of user.

Keywords: Colorization, A monochrome image, Landmark pixels extraction.

1 Introduction

A one-dimensional luminance scalar is replaced by a vector of a colorful multi-dimension for every pixel of a monochrome image, it is called as colorization. Obviously, it is under-constrained. Consequently, there has no only one result to colorization. In order to solve above problem, we should give some priori knowledge or set a reasonable constraint.

Some prior knowledge is considered to be given to the monochrome image. There are representative local colorization algorithms such as Welsh [1] which is a semi-automatic colorization algorithm. It transfers colors referring a colorful image to the greyscale image. However, it is no guarantee that the continuity of the colors in space. There are representative global colorization algorithms such as Levin [2] which is a colorization using optimization algorithm. Its basic idea: neighboring pixels in space and time that have similar intensities should have similar colors. User must draw some color scribbles to the monochrome image as some prior knowledge(Fig.3(a)). The indicated colors are propagated in both space and time to produce a fully colorized image. It may colorize the monochrome image in the context of not segmenting it to regions directly. It is an effective algorithm for a simple monochrome image.

However, Levin's method cannot effectively colorize monochrome images such as Fig.4(a) which is given as some scribbles at random by user. If we want to get the result of Fig3.(b), Levin's method must be repeated experiments for confirming the place of scribbles with Fig.4(a). It becomes a key point how to get some prior knowledge automatically. [5] obtained the distance of colors by repeating [2]'s method for extracting landmark pixels, while we obtained the distance of luminance by classification for extracting landmark pixels. Therefore, in our paper, it need not repeat experiments and automatically generates landmark pixels like scribbles. Landmark pixels are considered as the prior knowledge. We give some colors landmark pixels. Finally, colorize the monochrome image based on colorful landmark pixels according to requirements of user.

2 Algorithm

2.1 Generate Landmark Pixels

Landmark pixels are some representative pixels in an image. It costs much time if landmark pixels are extracted from an original image directly. Hence, degrade the monochrome image to low resolution image. The initial landmark pixels are extracted from the low resolution image using k-means. And then upgrade the resolution image, at the same time, we raise the number of landmark pixels. We repeat the above process until the result is the same as the resolution of the original image.

A monochrome image I_0 is given. We build a gaussian pyramid I_0, I_1, \dots, I_d , where I_0 is the input monochrome image of the original image and I_d is the coarsest level in the pyramid. We classify the coarsest level image I_d using information on the value of each pixel and position of each pixel. K clusters are obtained using k-means. The centroid of each cluster is considered as the initial landmark pixels. Let set of the initial landmark pixels be X_d . The mean value is substituted for the values of all pixels of each cluster. Let the image be Φ_d . The residue image is obtained by eq.(1) when $i = d$.

$$E_i = | I_i - \Phi_i | \quad (1)$$

We divide E_d into small windows like Fig.1. The size of each window is $h \times h$ pixels. Suppose the size of the input image is $M \times N$ pixels. We build a gaussian pyramid with a scale factor 2. Suppose that we set one landmark pixel to a window in the coarsest level image. c is the number of the landmark pixels. We will add at most $\frac{M}{2^k h} \times \frac{N}{2^k h}$ landmark pixels to small windows. Based on the idea, we can get h by eq.(2) and eq.(3). We set the same threshold to every small window. The value of the pixel should be memorized, if the mean value of pixel of the small window is larger than the threshold. The threshold is set at 30 based on experiments, if the number of landmark pixels is larger than 300. Otherwise, the threshold is set at 20. It is difficult to get the landmark pixels, if the threshold is too large. Otherwise, it will cost much time in order to extract many landmark pixels.

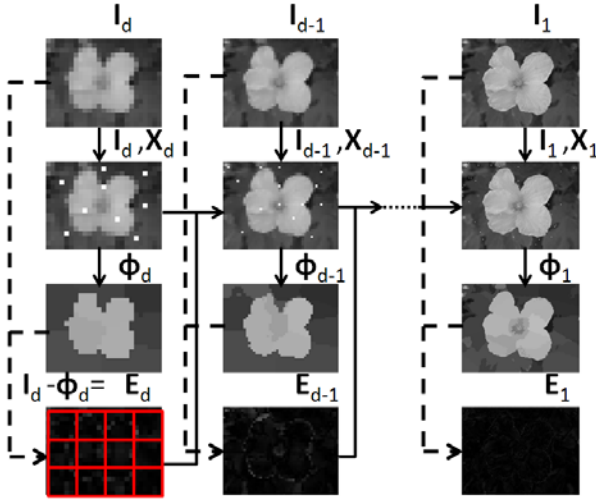


Fig. 1. Flowchart on process of generating landmark pixels

$$\sum_{k=0}^d \frac{M}{2^k h} \times \frac{N}{2^k h} = c \tag{2}$$

$$h = \sqrt{\frac{4MN}{3c} \left(1 - \frac{1}{4^{d+1}}\right)} \tag{3}$$

We can get the landmark pixels X_{d-1} from image E_d , Φ_{d-1} is obtained by segmenting I_{d-1} based on the set X_{d-1} of landmark pixels by K nearest neighbor. The residue image E_{d-1} can be obtained by eq.(1), when $k = d - 1$. In this way, until the landmark pixels X_0 are extracted from the image I_0 .

2.2 Classify Landmark Pixels

We should give every landmark pixel colorize the corresponding color. However, many landmark pixels are extracted from a monochrome image so that we are not able to colorize every landmark pixel. Fortunately, we find some landmark pixels should be set the same color. The same color landmark pixels would be colorized, if we only colorize a landmark pixel of the same color landmark pixels. According to this idea, we classify the landmark pixels to some clusters using ward's method. That is, the clusters of similarity have the small sum of squares while the clusters of difference have the large sum of squares in ward's method. We make the landmark pixels of the same cluster have the same color. We just colorize a landmark pixel of the same cluster, all landmark pixels obtain the same color in the cluster.

2.3 Colorize Landmark Pixels

How to colorize the monochrome image using the colored landmark pixels. We adopt Levin's method as her algorithm requires neither precise image segmentation, nor accurate region tracking. The basic idea of the algorithm: if neighboring pixels in space and time have similar intensities, they should have similar colors. That is to say, when the monochromatic luminance channel Y are similar, the chrominance channels U and V are similar. YUV color space is used in video [3]. In a word, it is a process to solve the solution of a quadratic cost function in sparse system of linear equations. The color scribbles are conditions of constraints in order to solve problem. In our paper, the color landmarks substitute for the color scribbles as conditions of constraints. The color landmarks more effective than the color scribbles without repeating experiments for confirming the place of scribbles.

2.4 Steps of Our Algorithm

Fig.2 shows the process of algorithm. It is carried out according to the following procedure.

1. Degrade an image to the low resolution image with downsampling.
2. Classify the low resolution image for initial landmark pixels.
3. Substitute the mean value of each cluster for the values of all pixels and obtain the image Φ .
4. Obtain the residue image E by $\|I - \Phi\|$.
5. Segment the residue image E with small windows so that the landmark pixels are added with these windows.
6. Classify the landmark pixels by ward's method.

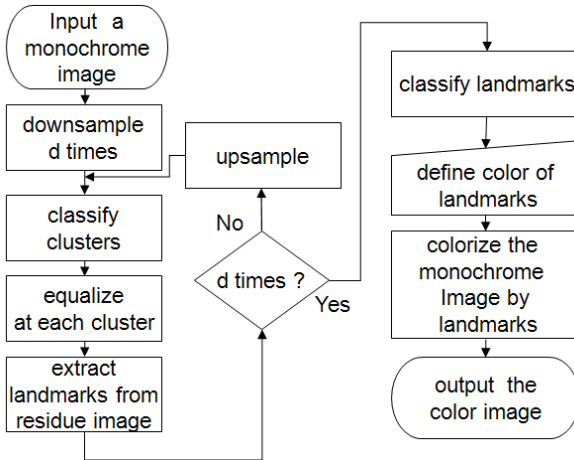


Fig. 2. Flowchart on process of our proposal

7. Define the landmark pixels of the same cluster to the same color.
 8. Colorize the monochrome image by the defined colored landmark pixels.
- Repeat from step 3 to step 5 until obtain the original image. After that, go ahead to step 6 until obtain a color image.

3 Experiments

Fig.3(a) and Fig.6(a) show the monochrome image with scribbles and its result with colorization from [2]. We draw some colored scribbles to the monochrome image freely like Fig.4(a) and Fig.7(a). We cannot obtain the result such as Fig.3(b) and Fig.6(b) while obtain Fig.4(b) and Fig.7(b). So we know it is not easy to get the result like Fig.3(b) and Fig.6(b). Experiments should be done until the result like Fig.3(b) and Fig.6(b) is obtained. Only by appropriately drawing the colored scribbles, Fig.3(b) and Fig.6(b) can be obtained. Our proposal does not consider the above problem for comparison. Our algorithm can generate landmark pixels automatically like Fig.5(a) and Fig.8(a). We just colorize one landmark pixel of each cluster and then can obtain the result like

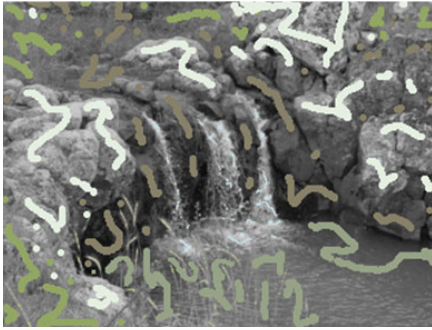


Fig. 3(a). The monochrome waterfall image with scribbled colors



Fig. 3(b). The result of the color waterfall image

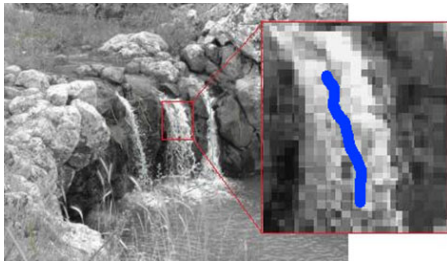


Fig. 4(a). The monochrome waterfall image with scribbled colors



Fig. 4(b). The result of the color waterfall image

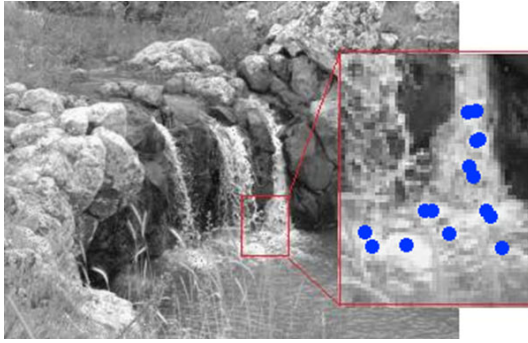


Fig. 5(a). The monochrome waterfall image with landmark pixels



Fig. 5(b). The result of the color waterfall image



Fig. 6(a). The monochrome child image with scribbled colors



Fig. 6(b). The result of the color child image



Fig. 7(a). The monochrome child image with scribbled colors

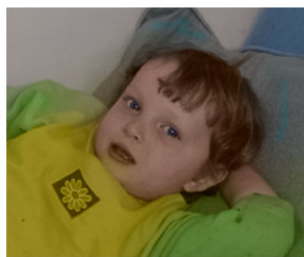


Fig. 7(b). The result of the color child image

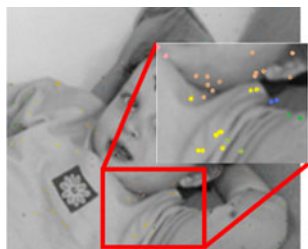


Fig. 8(a). The monochrome child image with landmark pixels



Fig. 8(b). The result of the color child image

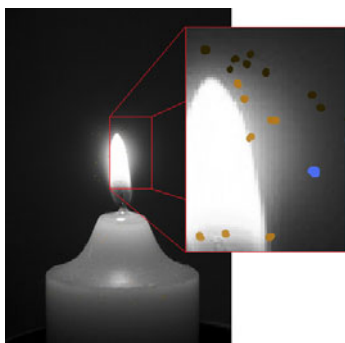


Fig. 9(a). The monochrome candle image with landmark pixels



Fig. 9(b). The result of the color candle image

Fig.5(b) and Fig.8(b). Some parameters of our proposal are given on Fig.5(b): the number of landmark pixels is $c=700$, threshold is $T=20$, the number of levels is $d=4$, the size of a small window is $h=13$, the number of clusters is $n=25$. On Fig.8(b): the number of landmark pixels is $c=300$, threshold is $T=20$, the number of levels is $d=5$, the size of a small window is $h=20$, the number of clusters is $n=100$. We carried out some experiments to other images. Their results are showed at Fig.9(b) and Fig.10(b).

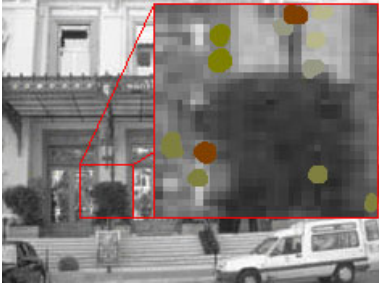


Fig. 10(a). The monochrome building image with landmark pixels



Fig. 10(b). The result of the color building image

4 Conclusion

In this paper, we present an effective colorization method using landmark pixels. Automatically generating landmark pixels is the advantage of the algorithm. However, we must define a color in the one landmark pixel of each cluster manually. Therefore, automatically defining a color in the one landmark pixel of each cluster is the subject of future research.

References

1. Welsh, T., Ashikhimin, M., Mueller, K.: Transferring color to greyscale images. *ACM Transactions on Graphics* 21(3), 277–280 (2002)
2. Levin, A., Lischinski, D., Weiss, Y.: MVA Conference: Colorization using optimization. In: *Proceedings of ACM SIGGRAPH 2004*, pp. 689–694 (2004)
3. Jack, K.: *Video demystified*, 3rd edn. Elsevier Science and Technology (2001)
4. Ward, J.H.: Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58, 236–244 (1963)
5. Huang, T.W., Chen, H.T.: Landmark-Based Sparse Color Representation for Color Transfer. In: *The 12th IEEE International Conference on Computer Vision*, pp.199–204 (2009)

Filtering-Based Noise Estimation for Denoising the Image Degraded by Gaussian Noise

Tuan-Anh Nguyen and Min-Cheol Hong

Video and Processing Laboratory, Information and Telecommunication Department,
Soongsil University, 156-743 Sangdo-Dong, Dongjak-Gu, Seoul, Korea
{anhnt, mhong}@ssu.ac.kr

Abstract. In this paper, a denoising algorithm for the Gaussian noise image using filtering-based estimation is presented. To adaptively deal with variety of the amount of noise corruption, the algorithm initially estimates the noise density from the degraded image. The standard deviation of the noise is computed from the different images between the noisy input and its' pre-filtered version. In addition, the modified Gaussian noise removal filter based on the local statistics such as local weighted mean, local weighted activity and local maximum is flexibly used to control the degree of noise suppression. Experimental results show the superior performance of the proposed filter algorithm compared to the other standard algorithms in terms of both subjective and objective evaluations.

Keywords: Local statistics, Gaussian filtering, noise estimation, Denoising, Gaussian noise.

1 Introduction

Noise having Gaussian-like distribution is very often encountered in acquired data. Gaussian noise is characterized by adding to each image pixel a value from a zero-mean Gaussian distribution. In the field of image processing and computer vision, noise removal while preserving image features such as edge, detail, and texture is a key problem [1].

In the past decades, there have been many attempts to construct digital filters which have the qualities of noise attenuation and detail preservation. For impulsive noise, the median filter and their modified approaches have been widely used due to their low computational cost benefits [1], [2], [3]. However, conventional noise reduction algorithms assume that the standard deviation of the Gaussian noise or noise level is known a priori, which is not valid in practical cases. Therefore, it is necessary to estimate the noise beforehand to apply these methods to some subsequent processes such as edge detection and object recognition for various applications.

In this paper, we propose a spatially adaptive denoising algorithm for image corrupted by Gaussian noise. Using local statistics to reflect a human visual system (HVS)[1], the noise level is obtained by filtering-based noise estimation, the noise detection function is defined to discriminate between true pixels and damaged one in the degraded image. In addition, under the assumption that an image is locally

Gaussian-distributed with different activity, a modified Gaussian filter is defined to remove corrupted components, where the parameters of the Gaussian filter are also determined by local statistics.

The rest of the paper is organized as follows. Section 2 addresses the noise variance estimation approaches. Constrains for noise estimation, the proposed noise detection and modified Gaussian filtering based on the local statistics are described in Section 3. In sections 4, the simulation results and performance comparisons will be presented to demonstrate the capability of the proposed algorithm. And finally section 5 reports conclusion.

2 Noise Variance Estimation Approaches

A common aim of image processing is to recover the original image x from the degraded image y that is contaminated by some noise model such as additive zero-mean white Gaussian noise n .

$$y = x + n. \quad (1)$$

Knowing the amount of noise is very important to allow other algorithms adaptively filtering images instead of using fixed thresholds. Many algorithms have been proposed in this field and usually the exact value of the noise variance σ_n^2 is required as a crucial filter parameter. However, the main difficulty intermixing of the statistics of original image x and the noise n due to (1). The separation of the two signals is not an easy task and it is well known that the noise variance of the sum of two independent signals is the sum of the variances of the two components.

Generally, noise estimation algorithms in the spatial domain are classified into two approaches: block-based and filtering-based (smoothing-based) [4]-[6].

In block-based methods [5], [6], images are tessellated into a number of $M \times N$ blocks. Standard deviations of intensity are computed for all the blocks and sorted. The block with the smallest standard deviation has the least change of intensity. The smaller the standard deviation is, the smoother the block. The intensity variation of a smooth block may be due to noise, in which the standard deviation of the block is close to that of the Gaussian noise added. The smallest standard deviation of a block of intensity is assumed to be equal to that of the additive white Gaussian noise. This algorithm is simple, but tends to overestimate the amount on noise for small noise cases. Noise can be underestimated in large noise cases. Also, the main difficulty of these methods is that their estimates may vary significantly depending on the input image and noise levels.

In filtering-based methods [4], a noisy input image is filtered by a low-pass filter to suppress the image structure. The standard deviation $\hat{\sigma}_n$ of the different image between the noisy input image and its filtered image is computed as illustrated in Fig. 1. This method especially yields good estimates for large noise cases [7].

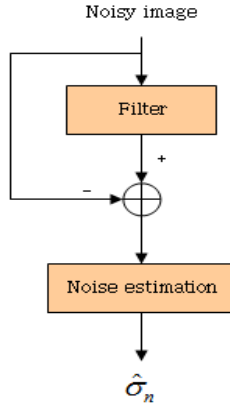


Fig. 1. Filtering-based noise estimation

3 Proposed Denoising Algorithm

The typical image degradation model at point (i, j) in a two dimensional coordinate is written as in (1). The problem at hand is to know exactly noise level on image in order to incorporate the most appropriate filter method. In this paper, a noise estimation algorithm bases on filtering using the local statistics is proposed. It is non-recursive and does not require the use of any kind of transforms. It shares the same characteristics in that each pixel is processed independently. Consequently, this approach has an obvious advantage when used in real-time digital image processing applications.

In addition, the local weighted mean, local weighted variance and local weighted max play an significant role to the denoising stage since its' advantages [8], [9], we use them as the key factors to establish the Modified Gaussian filter. For a pixel of the observed image y , the local weighted mean and the local weighted activity with the window of size $(2U + 1) \times (2V + 1)$ are defined as (2).

$$\begin{aligned} \mu_{i,j} &= \frac{\sum_m \sum_{n,(m,n) \in S_1} w_{m,n} \hat{x}_{i+m,j+n} + \sum_m \sum_{n,(m,n) \in S_2} w_{m,n} y_{i+m,j+n}}{\sum_{m=-U}^U \sum_{n=-V}^V w_{m,n}}, \\ \sigma_{i,j} &= \frac{\sum_m \sum_{n,(m,n) \in S_1} w_{m,n} |\hat{x}_{i+m,j+n} - \mu_{i,j}| + \sum_m \sum_{n,(m,n) \in S_2} w_{m,n} |y_{i+m,j+n} - \mu_{i,j}|}{\sum_{m=-U}^U \sum_{n=-V}^V w_{m,n}} \end{aligned} \tag{2}$$

In (2), $w_{m,n}$ denotes the weighting coefficient at the point (m, n) within the window, and $|\bullet|$ represents the absolute operator. Also, $\hat{x}_{i+m,j+n}$ denotes the value of the reconstructed pixel at the point $(i + m, j + n)$ belonging to S_1 , where S_1 and S_2 represent the causal region (dark region) and the non-causal region (white region) with respect to the point (i, j) in progressive scanning order, and the intersection between S_1 and S_2 is null, as shown in Fig. 2.

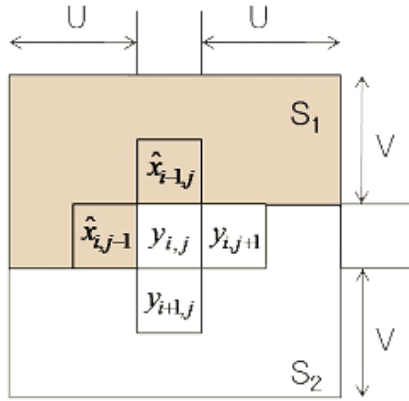


Fig. 2. Local window for determining weighted local information

Also, the local maximum of the observed image is

$$y_{\max,i,j} = \max_{(p,q) \in S} y_{p,q}. \tag{3}$$

where S is the support region to determine the local maximum about the point (i, j) . In this work, S is the same as the analysis window used for local weighted mean and local weighted activity ($S_1 \cup S_2 = S$). In order to effectively obtain the better pre-filtered image, we define the noise detection, which can discriminate noisy and noise-free pixels prior to applying the noise removal filter. Under the assumption that an image is locally Gaussian-distributed with local smoothing constraint, a noise detection function as in (4)

$$flag_{i,j} = \begin{cases} 1 & \text{if } y_{i,j} > \mu_{i,j} + B_{i,j} \\ & \text{or } y_{i,j} < \mu_{i,j} - B_{i,j} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$B_{i,j} = k \times \frac{\sigma_{i,j}}{y_{\max,i,j}},$$

In (4), k represents a constant. This equation means that the value of the flag is equal to 1 when a pixel is detected as a corrupted one. It is clear how the local statistics affects the detection function. Assume that a same noise is added to both the flat area and the high activity region with the same local maximum. In this case, $B_{i,j}$ of the flat area is relatively smaller than that of the high activity region. The smaller $B_{i,j}$ represents tighter bounds for the flat region, so that small variations of the flat can be detected as corrupted pixels. On the other hand, a higher activity region leads

to looser bounds and small variations of the high activity area are not detected as uncorrupted pixels, leading to the preservation of important features. This is in agreement with the noise masking property of the HVS [1].

According to the above condition, when a pixel is detected as the corrupted one with lower local weighted activity, strong filtering (over-smoothing) process is required, while weak filtering for the corrupted pixels with higher activity. Gaussian filter is very useful to control the degree of the smoothness of the reconstructed image by using the local activity.

Using the local statistics in (2), we propose a modified Gaussian filter such as

$$h_{i,j} = \frac{1}{Z_{i,j}} \exp\left(-T \frac{\sigma_{i,j}^2(i^2 + j^2)}{\sqrt{\mu_{i,j} + 1}}\right). \quad (5)$$

where $Z_{i,j}$ and T denote the normalizing constant and a tuning parameter, respectively. Also, the support region of the Gaussian filter in (5) is the same as that of the analysis window in (3). Then, the reconstructed pixel can be written as (6). For the corrupted pixels with small local activity (flat region), (6) leads to strong filtering since the filter coefficients within the support region have similar values. On the other hand, it results in weak filtering for the corrupted pixels with large local activity.

$$\hat{x}_{i,j} = \begin{cases} \frac{\sum_m \sum_{n,(m,n) \in S_1} h_{m,n} \hat{x}_{i+m,j+n} + \sum_m \sum_{n,(m,n) \in S_2} h_{m,n} y_{i+m,j+n}}{\sum_{m=-U}^U \sum_{n=-V}^V h_{m,n}} & \text{if } flag_{i,j} = 1 \\ y_{i,j} & \text{otherwise} \end{cases} \quad (6)$$

To incorporate the most appropriate filtering method, the noise variance levels on the input degraded image are estimated as in Fig. 3, in which as the estimated noise less than the value α , the weighting coefficients are the diagonal element within the $(2U+1) \times (2V+1)$ support window to avoid the over-smoothness problem. In this case, it means that the noise level is relatively low (greater than 20 dB and 30 dB). Hence, it is better to use only 5 pixels in the *cross* region of the 3×3 support window to calculate the local weighted mean $\mu_{i,j}$, local weighted variance $\sigma_{i,j}$ and the output pixel $\hat{x}_{i,j}$. On the other hand, as the estimated noise greater than the α , the input image is corrupted seriously (less than 10 dB). It is, therefore, necessary to incorporate more pixels for the processing. The uniform weighting coefficients and all pixels within the $(2U+1) \times (2V+1)$ support window are used to calculate the local information.

The novelty of the proposed algorithm is that it has the capability to effectively remove the noise components using the local statistics and estimate the image's noise density for flexibly treats different amount of corruptions.

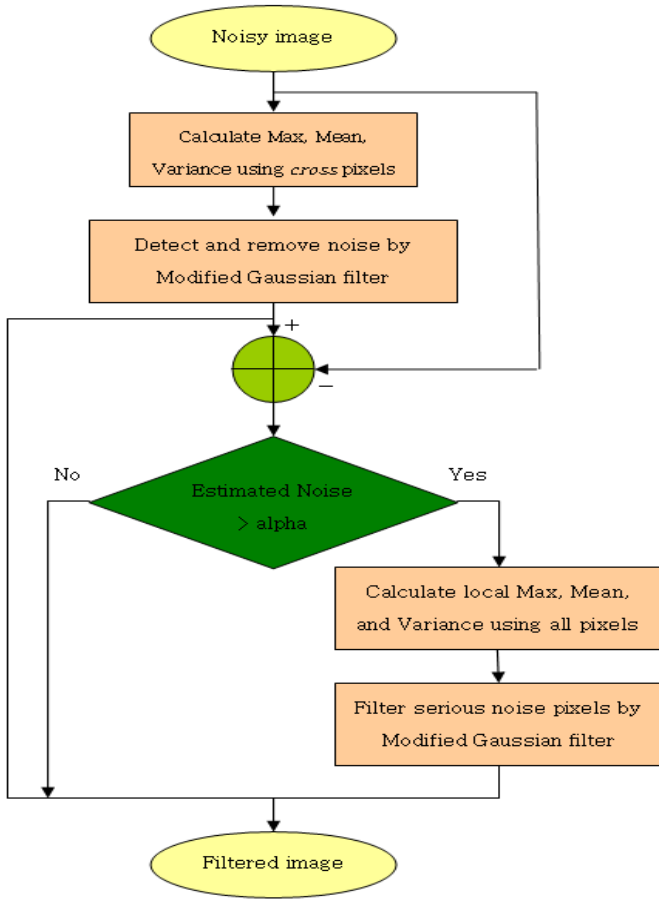


Fig. 3. Entire procedure of proposed algorithm

4 Experimental Results and Performance Comparisons

The proposed algorithm was tested with various Gaussian noise images for various SNRs such as “Lena”, “Cameraman”, “Bird”, “Goldhill” images and compared them with pixel-wise median absolute difference (PWMAD)[10], rank-order criterion filter (ROC)[11], switching-based adaptive weighted mean (SAWM)[12] and bilateral filter (BF)[13].

In order to evaluate the performance of the noise detection algorithm, the following noise detection fidelity (D_F) was used. It is

$$D_F = \left(1 - \frac{F_p + M_p}{T_p} \right) \times 100. \tag{7}$$

In (7), F_p , M_p , and T_p represent the number of detection “fault” pixels, the number of detection “missing” pixels, and the number of total pixels in an image, where “fault” means that an uncorrupted pixel is detected as a corrupted one. On the other hand, “missing” denotes that a corrupted pixel is considered as an uncorrupted one. In our work, $alpha = 4$, $U = 1$, $V = 1$ and the uniform weighting coefficients in (2) are used. For evaluating the performance of noise filtering, we use peak signal-to-noise ratio (PSNR) and universal image quality index (UIQI) which is a new method for image assessment [14].

Table 1. Performance comparisons of Lena image

Noise	Method	D_F	PSNR	UIQI
10dB	PWMAD	82.04	27.77	0.639
	ROC	81.91	27.61	0.635
	SAWM	84.96	27.84	0.632
	BF	N/A	29.43	0.698
	Proposed	94.98	29.45	0.718
20dB	PWMAD	71.83	30.64	0.797
	ROC	70.43	30.18	0.796
	SAWM	84.86	30.95	0.800
	BF	N/A	32.28	0.830
	Proposed	89.61	32.36	0.851
30dB	PWMAD	67.37	31.24	0.852
	ROC	55.84	31.66	0.854
	SAWM	70.03	31.9	0.857
	BF	N/A	32.63	0.860
	Proposed	70.03	33.09	0.914

Table 2. Performance comparisons of Cameraman image

Noise	Method	D_F	PSNR	UIQI
10dB	PWMAD	83.94	25.08	0.369
	ROC	82.67	25.06	0.368
	SAWM	90.14	25.93	0.389
	BF	N/A	27.93	0.441
	Proposed	95.65	28.11	0.445
20dB	PWMAD	83.85	27.06	0.516
	ROC	78.65	27.34	0.515
	SAWM	86.02	27.38	0.522
	BF	N/A	31.22	0.583
	Proposed	93.62	29.77	0.612
30dB	PWMAD	70.32	27.32	0.644
	ROC	69.33	27.76	0.656
	SAWM	70.95	27.7	0.649
	BF	N/A	31.62	0.678
	Proposed	79.04	30.31	0.784

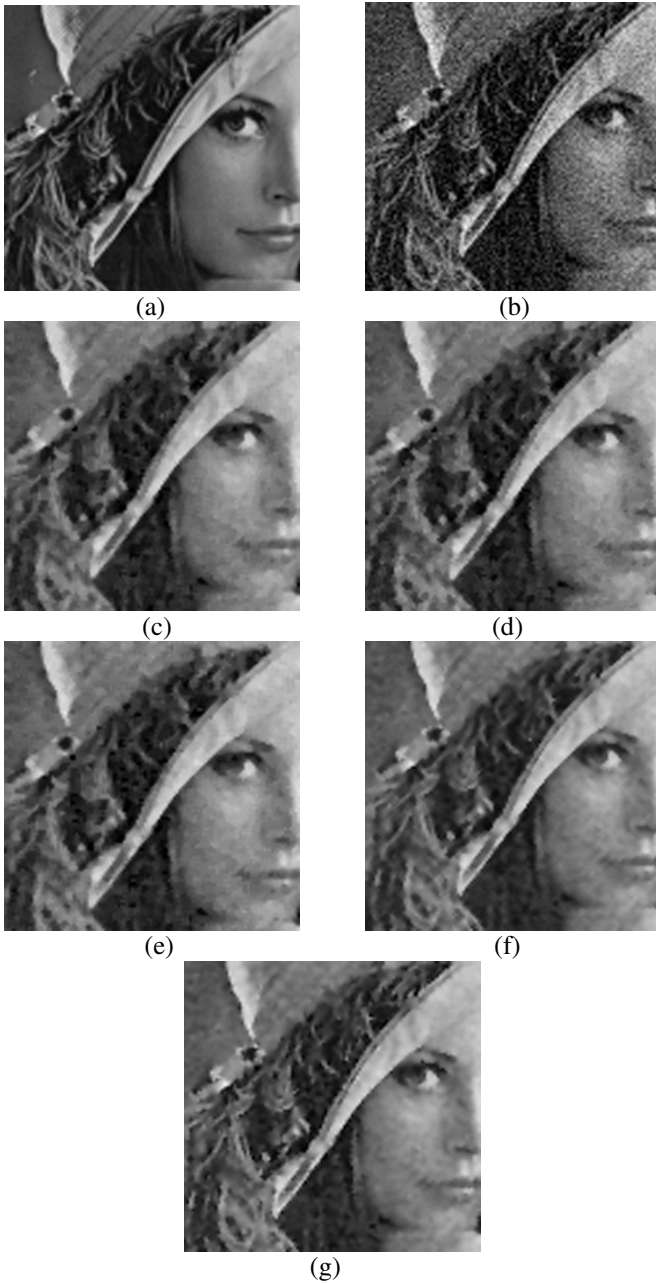


Fig. 4. Experimental results of Lena image: (a) enlarged original image, (b) enlarged degraded image with 10dB Gaussian noise, (c) corresponding reconstructed image with PWMAD, (d) corresponding reconstructed image with ROC, (e) corresponding reconstructed image with SAWM, (f) corresponding reconstructed image with BF, (g) corresponding reconstructed image with proposed algorithm

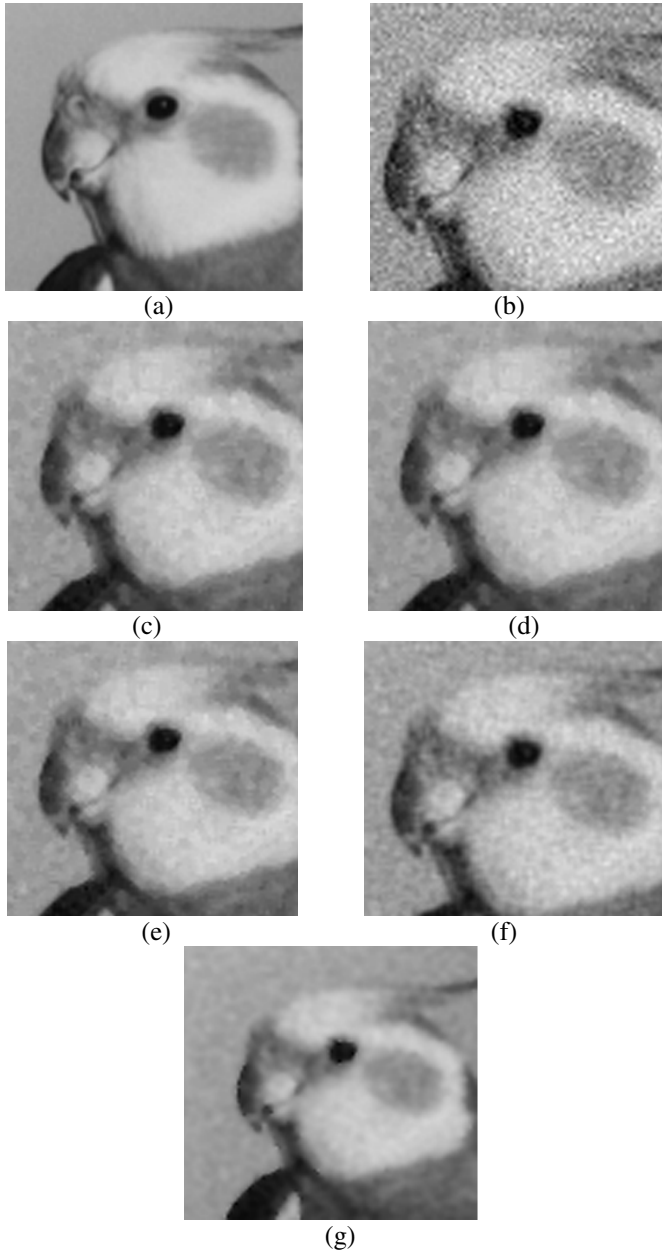


Fig. 5. Experimental results of Bird image: (a) enlarged original image, (b) enlarged degraded image with 10dB Gaussian noise, (c) corresponding reconstructed image with PWMAD, (d) corresponding reconstructed image with ROC, (e) corresponding reconstructed image with SAWM, (f) corresponding reconstructed image with BF, (g) corresponding reconstructed image with proposed algorithm

As shown in Fig. 4 and Fig. 5, PWMAD, ROC, SAWM lead to over-blurred reconstructed results. We observed that the degree of the smoothness is more serious as the additive noise is smaller since the degree of the additive noise is not considered in the filtering process of the methods. On the other hand, BF and proposed algorithm lead to the effective noise removal with preservation of the important features.

Table 1 and Table 2 summarize the performance comparisons with respect to D_F , PSNR and UIQI. They show that D_F of ours outperforms the other approaches for all cases (BF does not require noise detection procedure). In addition, it was observed that BF and proposed algorithm are the most competitive among the above approaches in terms of PSNR and UIQI. However, the gain is getting smaller as the images are seriously contaminated.

5 Conclusions

In this paper, a denoising algorithm for the Gaussian noise image using filtering-based estimation is presented. This shows an efficient noise denoising algorithm that takes into account the filtering-based noise estimation algorithm and noise detection that leads to objectively and subjectively satisfactory results without prior information about the noise by incorporating the local statistics. Currently, we are improving the noise estimate stage to obtain a more sophisticated formulation can be derived and better performance.

Acknowledgments. This work was supported by the Korea Science and Engineering Foundation (KOSEF) grant fund by the Korea Government (MEST) (No. 2011-0000148) and the Ministry of Knowledge Economy, Korea under the Information Technology Research Center support program supervised by the National IT Industry Promotion Agency [NIPA-2011-(C1090-1111-0003)].

References

1. Arce, G.R.: Nonlinear signal processing: A Statistical approach. John Wiley and Sons Inc. (2004)
2. Nodes, T.A., Gallagher, N.C.: Median filters: some modifications and their properties. *IEEE Trans. Acoustics, Speech and Signal process.* 30(5), 739–746 (1982)
3. Bednar, J.B., Watt, T.K.: Alpha-trimmed means and their relationship to median filter. *IEEE Trans. Acoustics, Speech and Signal Process.* 32(1), 145–153 (1984)
4. Olsen, S.I.: Noise Variance Estimation in Images: An evaluation, *Computer Vision Graphics Image Processing. Graphic Models and Image Processing* 55(4), 319–323 (1993)
5. Lee, J.S., Hoppel, K.: Noise modeling and estimation of remotely-sensed image. In: *International Conference on Geoscience and Remote Sensing, Vancouver, Canada, vol. 2*, pp. 1005–1008 (1989)
6. Shin, D.H., Park, R.H., Yang, S.J.: Block-based noise estimation using adaptive Gaussian filtering. *IEEE Trans. on Consumer Electronics* 51(1) (2005)
7. Rank, K., Lendl, M., Unbehauen, R.: Estimation of image noise variance. *IEEE Proc. Vision Image Signal Process.* 146, 8–84 (1999)

8. Lee, J.S.: Refined filtering of image noise using local statistics. *Computer Vision, Graphics and Image processing* 15, 380–389 (1989)
9. Mastin, G.A.: Adaptive filters for Digital noise smoothing, An evaluation. *Computer vision, Graphics and Image processing* 31, 103–121 (1985)
10. Crnojevic, V., Senk, V., Trpovski, Z.: Advanced impulse detection based on pixel-wise MAD. *IEEE Signal Process. Letters* 11(7), 589–592 (2004)
11. Aizenberg, I., Butakoff, C.: Effective impulse detector based on rank-order criteria. *IEEE Signal Process. Letters* 11(3), 363–366 (2004)
12. Zhang, X., Xiong, Y.: Impulse noise removal using directional differences based noise detector and adaptive weighted mean filter. *IEEE Signal Process. Letters* 16(4), 295–298 (2009)
13. Elad, M.: On the origin of the bilateral filter and ways to improve it. *IEEE Trans. Image Process.* 11(10), 1141–1151 (2002)
14. Wang, Z., Bovik, A.C.: A universal image quality index. *IEEE Signal Processing Letters* 9(3), 81–84 (2002)

Combining Mendonça-Cipolla Self-calibration and Scene Constraints

Adlane Habed¹, Tarik Elamsy², and Boubakeur Boufama²

¹ Université de Bourgogne
Laboratoire Le2i, UMR CNRS 5158
BP16, Route des plaines de l'Yonne
89010 Auxerre Cedex, France
`adlane.habed@u-bourgogne.fr`

² University of Windsor
School of Computer Science
401 Sunset Avenue
Windsor, ON. N9B 3P4, Canada
`{elamsy,boufama}@uwindsor.ca`

Abstract. In this paper, we propose a method that combines plane parallelism and the Mendonça/Cipolla self-calibration constraints. In our method each pair of images is treated independently and can therefore use a different pair of parallel planes not necessarily visible in the other views. While, for each pair of images, constraints on the singular values of the essential matrix provide two algebraic constraints on the intrinsic parameters, those we derive from plane parallelism have the advantage of providing two additional ones making the calibration of a no-skew camera possible from two images only.

Keywords: self-calibration, 3D reconstruction, scene constraints.

1 Introduction

The problem of self-calibrating a camera, i.e. retrieving its intrinsic parameters solely from point correspondences across images, has been extensively addressed in the literature. It is both highly desirable and important that camera self-calibration methods not only should they exhibit nice convergence, stability, robustness and accuracy properties but also have the ability to easily incorporate scene constraints such as parallelism and orthogonality whenever these are available. For instance, incorporating scene constraints in the self-calibration process may contribute in reducing the number of required images, obtaining more accurate 3D reconstructions and parameters as well as overcoming degeneracy issues generally due to special camera motions.

Vanishing points, possibly in conjunction with other constraints (such as the modulus constraints [8]), are commonly used to help obtaining an estimate of the plane at infinity and hence recover the intrinsic parameters. Ideally, however,

scene constraints need to be incorporated in the process of calculating the intrinsic parameters as to reduce degeneracy situations. These parameters are generally calculated through the recovery of the projective representations of either the so-called Image of the Absolute Conic (or its dual the DIAC), the Absolute Plane Quadric or the Absolute Line Quadric, all of which embed the internal geometry of the camera (the reader may refer to [3] and the reference therein for more details on the subject). The geometry of these conics and quadrics is well suited for expressing parallel and/or orthogonal directional constraints in a way these can be easily combined with self-calibration constraints [2,6,5]. Quadric-based self-calibration methods require the prior calculation of camera matrices that are consistent with the same projective or quasi-affine [3] 3D reconstruction of the scene. These methods are known to perform better than their conic-based counterpart which are simpler, as they only require the epipolar geometry relating pairs images, but exhibit poor numerical stability and high sensitivity to both noise and initialization.

Also relying on the sole calculation of the epipolar geometry between pairs of views, Mendonça and Cipolla [7] have proposed an even simpler, yet effective, method which, unlike conic-based methods, exhibits remarkable convergence properties and recovers camera parameters with a fair accuracy [1]. Their method allows to directly estimate the camera parameters through constraints on the singular values of the essential matrix relating each pair of views. Despite the simplicity and the good performance of their self-calibration method, there has been, to our knowledge, no attempt to combine their constraints on the essential matrix and those from the scene.

In this paper, we propose a method that nicely combines plane parallelism and the Mendonça/Cipolla self-calibration constraints without explicitly referring to the geometry of neither the Image of the Absolute Conic nor its Dual. In our method, each pair of images is treated independently and can therefore use a different pair of parallel planes not necessarily visible in the other views. While, for each pair of images, constraints on the singular values of the essential matrix provide two constraints on the intrinsic parameters, those we derive from plane parallelism provide two additional ones making the calibration of a no-skew camera possible when using two images only. In addition, our experiments show that the proposed method significantly improves the quality of the estimated 3D reconstruction, that of the intrinsic parameters and the convergence of the algorithm in comparison with Mendonça/Cipolla self-calibration. Our results also show that this new method outperforms, in terms of accuracy of the 3D reconstruction, the recent stratified method presented in [2] which also exploits plane parallelism.

Our paper is organized as follows. A review of the Mendonça-Cipolla self-calibration method is presented in Section 2. Our new constraints, which are due to plane parallelism, are derived in Section 3. A method for combining plane parallelism and the Mendonça and Cipolla self-calibration constraints is described in Section 4. The results of our experiments are detailed in Section 5. Section 6 concludes our work.

2 Direct Self-calibration Constraints

Assuming the cameras satisfy the pinhole model, it is well-known that two corresponding points p_i and p_j in two images i and j are related by the epipolar constraint: $\mathbf{p}_j^T \mathbf{F}_{ij} \mathbf{p}_i = 0$. \mathbf{F}_{ij} is a rank-2 3×3 matrix known as the fundamental matrix while \mathbf{p}_i and \mathbf{p}_j are the homogeneous pixel coordinates of the projections, in the two images, of the same scene point. The fundamental matrix is closely related to the so-called essential matrix

$$\mathbf{E}_{ij} = [\mathbf{t}_{ji}]_{\times} \mathbf{R}_{ij} \quad (1)$$

which embeds the translational \mathbf{t}_{ji} and the rotational \mathbf{R}_{ij} components of the rigid motion between the two cameras. Unlike the fundamental matrix, the essential matrix cannot be calculated from point correspondences across images unless the intrinsic parameters of both cameras are known [3]. These matrices are indeed related through

$$\mathbf{E}_{ij} \simeq \mathbf{A}_j^T \mathbf{F}_{ij} \mathbf{A}_i \quad (2)$$

where the \mathbf{A} matrices are 3×3 upper-triangular and parameterized in terms of the pixel-valued focal lengths, f_u and f_v , along the main axes of the image, the pixel coordinates (u_0, v_0) of the principal point, and the skew factor s of the camera under consideration; that is,

$$\mathbf{A} = \begin{bmatrix} f_u & s & u_0 \\ 0 & f_v & v_0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (3)$$

Mendonça and Cipolla's self-calibration method [7] is based on the property of the essential matrix having two equal singular values while the third one is zero. This property has been proved by Huang and Faugeras [4] who have also shown that this is a necessary and sufficient condition for a matrix to be factored into the product of an orthogonal and a skew-symmetric matrix. Moreover, it has been demonstrated that the equality of the singular values imposes two algebraic constraints on the essential matrix. Assuming the cameras have the same intrinsic parameters (e.g. a moving camera with constant parameters), the self-calibration method proposed in [7] consists in minimizing over $n \geq 3$ views the cost function

$$\mathcal{C}(\mathbf{A}) = \sum_{ij}^n \frac{w_{ij}}{\sum_{kl}^n w_{kl}} \frac{\sigma_{ij}^+ - \sigma_{ij}^-}{\sigma_{ij}^-} \quad (4)$$

where σ_{ij}^+ and σ_{ij}^- are respectively the largest and the smallest non-zero singular values of $\mathbf{A}^T \mathbf{F}_{ij} \mathbf{A}$ and w_{ij} are weights that are inversely proportional to a measurement of the quality of the calculated fundamental matrix.

3 Scene Constraints

We consider in the following a pair of parallel planes Π and Φ observed by two cameras (from a sequence) whose fundamental matrix \mathbf{F}_{ij} and epipole \mathbf{e}_{ji}

$(F_{ij}^T \mathbf{e}_{ji} = 0)$ have already been recovered [3]. Being parallel, the two planes meet in a line on the plane at infinity. As a consequence, the 3D homogeneous coordinate vectors of our planes and that of the plane at infinity are linearly dependent. Such relationship is independent from the choice of the projective reference frame in which the coordinates of these planes are expressed. However, each pair of cameras, when considered independently from the others, allows to reconstruct the scene in a frame that is different from any frame chosen for some other pair of cameras. In order to emphasize this fact, we will denote by Π_{ij} , Φ_{ij} and ∞_{ij} the homogeneous coordinate vectors of Π , Φ and the plane at infinity, respectively, expressed in a projective frame chosen only for the pair of cameras i and j under consideration. Moreover, because plane coordinates are homogeneous and assuming none of the considered planes passes through the origin of the reference frame, we choose our plane coordinates scaled such that

$$\Pi_{ij}^T = (\pi_{ij}^T \ 1), \quad \Phi_{ij}^T = (\phi_{ij}^T \ 1) \text{ and } \infty_{ij}^T = (\alpha_{ij}^T \ 1). \quad (5)$$

In general, the linear dependency of these three planes is expressed through

$$\infty_{ij} = \alpha_{ij}\Pi_{ij} + \beta_{ij}\Phi_{ij} \quad (6)$$

for some non-zero scalars α_{ij} and β_{ij} . However, for the choice of coordinates we have made in (5), these scalars must also verify $\alpha_{ij} + \beta_{ij} = 1$. This allows to eliminate one of the scalars and thus further simplify the relationship between the directional components of the three planes; that is,

$$\alpha_{ij} = \alpha_{ij}\pi_{ij} + (1 - \alpha_{ij})\phi_{ij}. \quad (7)$$

Since we are free to choose the projective frame in which to express the scene and cameras, we might as well choose it such that the reference frame is attached to camera i and the projection matrices of the two cameras given by $M_i = [I \ | \ 0]$ and $M_j = [[\mathbf{e}_{ji}]_{\times} F_{ij} \ | \ \mathbf{e}_{ji}]$. This choice of matrices is often considered in the literature as a "starting" projective reconstruction [3] since it is fairly simple to obtain. In particular, it allows us to use only the epipolar geometry of the pair of cameras in order to retrieve the vectors π_{ij} and ϕ_{ij} in the frame under consideration. For instance, if \mathbf{p}_i and \mathbf{p}_j are the homogeneous pixel coordinates of the projections in images i and j of the same 3D point P on a plane Π , our choice of M_i indicates that $P \simeq (\mathbf{p}_i \ \lambda_i)$ for some value of the projective depth λ_i . Because P lies on Π , the projective depth is given by $\lambda_i = \pi_{ij}^T \mathbf{p}_i$. Projecting P on image j via the projection matrix M_j leads to the relationship $\mathbf{p}_j \simeq ([\mathbf{e}_{ji}]_{\times} F_{ij} + \mathbf{e}_{ji} \pi_{ij}^T) \mathbf{p}_i$. Since the latter relationship is true for any plane, the inter-image homographies induced by Π and Φ are given by

$$H_{ij\Pi} \simeq [\mathbf{e}_{ji}]_{\times} F_{ij} + \mathbf{e}_{ji} \pi_{ij}^T \text{ and } H_{ij\Phi} \simeq [\mathbf{e}_{ji}]_{\times} F_{ij} + \mathbf{e}_{ji} \phi_{ij}^T \quad (8)$$

which allow the calculation of π_{ij} and ϕ_{ij} from point correspondences assuming points on each plane are identified and their projections in the two matched images. Similarly, the inter-image homography of the plane at infinity will also be on the form

$$H_{ij\infty} \simeq [\mathbf{e}_{ji}]_{\times} F_{ij} + \mathbf{e}_{ji} \alpha_{ij}^T. \quad (9)$$

where α_{ij} remains unknown. Using the relationship (7) that is due to plane parallelism, we can however reduce $H_{ij\infty}$ to the 1D family of matrices:

$$H_{ij\infty} \simeq [e_{ji}]_{\times} F_{ij} + \alpha_{ij} e_{ji} \pi_{ij}^T + (1 - \alpha_{ij}) e_{ji} \phi_{ij}^T. \quad (10)$$

The product $H_{ij\infty} [\pi_{ij} - \phi_{ij}]_{\times}$ eliminates the unknown scalar α_{ij} from the above expression leading after simplification to

$$H_{ij\infty} [\pi_{ij} - \phi_{ij}]_{\times} \simeq [e_{ji}]_{\times} F_{ij} [\pi_{ij} - \phi_{ij}]_{\times} - e_{ji} \pi_{ij}^T [\phi_{ij}]_{\times}. \quad (11)$$

Moreover, because $H_{ij\infty}$ can be factored into $H_{ij\infty} = A_j R_{ij} A_i^{-1}$, one can easily deduce that, just like the essential matrix, $A_j^{-1} H_{ij\infty} [\pi_{ij} - \phi_{ij}]_{\times} A_i^{-T}$ can be factored into the product $R_{ij}[q]_{\times}$ of the rotation matrix R_{ij} and a skew-symmetric matrix $[q]_{\times} = A_i^{-1} [\pi_{ij} - \phi_{ij}]_{\times} A_i^{-T}$. As a result, the matrix $D_{ij\pi\phi}$ such that

$$D_{ij\pi\phi} = A_j^{-1} ([e_{ji}]_{\times} F_{ij} [\pi_{ij} - \phi_{ij}]_{\times} - e_{ji} \pi_{ij}^T [\phi_{ij}]_{\times}) A_i^{-T} \quad (12)$$

must have two equal singular values while the third one is zero. The existence of the zero singular value is by construction of the matrix $D_{ij\pi\phi}$. However, the equality of the other two singular values provides two additional algebraic constraints on the intrinsic parameters of the camera.

4 Constraints Combination

In general, when scene and self-calibration constraints are combined, the problem is cast into a constrained nonlinear optimization one. This is, for instance, the case in [5] where self-calibration and orthogonal planes constraints are combined. However, because the scene and all the cameras are reconstructed in the same frame, each pair of such planes provides only one constraint regardless of the length of the sequence. In the present paper, the situation is different since, in the absence of common frame for all cameras, the fact that two planes are parallel must be taken into account by every pair of images in which those planes are visible. Moreover, for each pair of images the constraints on $D_{ij\pi\phi}$ are as important as those on the essential matrix since they introduce the same number of algebraic constraints. As a consequence, we have chosen to cast the problem of combining the constraints on the essential matrix E_{ij} and those on $D_{ij\pi\phi}$ in an unconstrained optimization procedure. For $n \geq 2$ images captured by a moving camera (constant parameters), we propose the following cost function

$$\mathcal{M}(A) = \sum_{ij}^n \frac{w_{ij}}{\sum_{kl}^n w_{kl}} \frac{\sigma_{ij}^+ - \sigma_{ij}^-}{\sigma_{ij}^-} + \sum_{(\Pi, \Phi) \in \mathcal{S}} \gamma_{\Pi\Phi} \sum_{ij}^n \frac{v_{ij\Pi\Phi}}{\sum_{kl}^n v_{kl\Pi\Phi}} \frac{\delta_{ij\Pi\Phi}^+ - \delta_{ij\Pi\Phi}^-}{\delta_{ij\Pi\Phi}^-} \quad (13)$$

where \mathcal{S} is the set of all pairs of identified parallel planes in the scene. The weights $v_{ij\Pi\Phi}$ are inversely proportional to the quality of the homography matrices (8) of the planes under consideration if the latter are visible in images i and j . These weights are otherwise null. While w_{ij} , σ_{ij}^+ and σ_{ij}^- are as in (4), $\delta_{ij\Pi\Phi}^+$ and

$\delta_{ij\Pi\Phi}^-$ are respectively the largest and the smallest non-zero singular values of $D_{ij\Pi\Phi}$. Furthermore, $\gamma_{\Pi\Phi}$ is a parameter chosen to reflect the confidence one has in the constraint that two planes are parallel. In all our experiments, we have set $\gamma_{\Pi\Phi} = 1$. Note that with two images and only one pair of parallel planes, 4 constraints are imposed on the intrinsic parameters of the camera which suffices in theory to calibrate a camera with no skew.

5 Experiments

We provide here some of the results obtained through extensive experiments using simulated data and two examples with real images.

5.1 Simulated Data

The goal of our experiments with simulated data was twofold: (1) assess the quality of the estimated 3D structure and that of the intrinsic parameters using both Mendonça/Cipolla's self-calibration and our method with parallel planes; (2) compare the results obtained with our method against those obtained using the stratified method [2]. Note that this stratified method uses plane parallelism to estimate the position of the plane at infinity and, unlike our method, it does not directly relate parallelism constraints to the intrinsic parameters. Using [2], an additional step for estimating the intrinsic parameters and another one for refining them have been used in our experiments.

All our simulations were carried out for various lengths of the image sequence and levels of noise added to image coordinates. For each sequence length and noise level, we have tested each method on 200 independent, randomly generated, sets of scenes and cameras. Each scene consisted of a randomly generated pair of parallel planes. 50 scene points scattered in a disc of unit-radius have been randomly generated for each plane. These points have been then projected onto a set of images each with a size of 512×512 pixels. The first plane was generated such that its disc was centered in the origin of the scene's reference frame. The second plane was generated parallel to the first one and at a mean distance of 0.5 units (0.25 standard deviation) from it. The cameras were generated at a mean distance of 2.5 units with 0.25 standard deviation from the center of the scene and roughly oriented towards the origin of the scene's reference frame. Camera parameters $f_u = f_v = 800$, $s = 0$ and $u_0 = v_0 = 256$ were kept constant throughout the sequence. Zero-mean Gaussian noise with standard deviation in the range 0 to 2 pixels (with a step of 0.25 pixel) was added to the pixel coordinates. The intrinsic parameters estimated by each method along with the ground truth overall scale of the scene have been used to recover the 3D Euclidean structure by triangulation. The 3D Root-Mean-Square (RMS) error of the reconstructed points was calculated. Only simulations leading to a 3D RMS error < 0.65 were considered as successful. When using the stratified [2], the DIAC was also required to pass the positive-definiteness test. For each noise level and sequence length, the average 3D RMS error and the relative RMS

error on the estimated intrinsic parameters were calculated over the successful trials out of the 200 independent runs. The skew factor s was always assumed to be known. Following this account, we have carried out two independent sets of experiments:

- the first set involved our method with parallel planes versus Mendonça/Cipolla’s self-calibration method. For each simulated scene and cameras, optimization was started from the same initial guess of the parameters using both the scene-constraints-free objective (4) and the one incorporating plane parallelism (13). At each trial, we have considered two initial values of the intrinsic parameters to start optimization from:
 - (a) the true noise-free values of the intrinsic parameters,
 - (b) randomly chosen parameters in which the values of f_u and f_v were picked in the range 1 to 2000 while those of the coordinates u_0 and v_0 were taken in a 200×200 pixels area centered at image coordinates (256, 256).
- the second set of experiments involved the method presented in this paper versus the stratified method [2]. Optimization of (13) was randomly started as described in (b) for the first set of experiments. The method in [2] does not require initialization: it proceeds by identifying the plane at infinity through solving sets of quartic equations.

The results obtained using our method with one pair of parallel planes and two images (for various levels of noise) are summarized in Fig. 1. The average 3D RMS errors, Fig. 1-(a), and the relative RMS errors (in %) on the intrinsic parameters, Fig. 1-(a), were calculated after initializing our objective function (13) from the true noise-free intrinsic parameters. These errors have been calculated over the successful trials (i.e. 3D RMS error < 0.65) whose rate is given in Fig. 1-(c). On the other hand Fig. 1-(d) provides the convergence rate (convergence to the correct parameters) of the optimization when the latter is randomly started. These results show that the errors obtained are fair considering only two images have been used. However, a good initial guess is definitely required in such minimal situation even in the absence (or low levels) of noise.

The results of similar experiments with 3 images, using both our method with one pair of parallel planes and the Mendonça-Cipolla’s, are summarized in Fig. 2. It is clear from these results that incorporating plane parallelism constraints plays a very important role in improving not only the quality of the reconstruction (Fig. 2-(a)) and the parameters (Fig. 2-(b) for our method and Fig. 2-(c) for Mendonça and Cipolla’s) but also the convergence of the algorithm. It can be seen in Fig. 2-(e) that our method has converged (from random start) in 85% of the trials with 2 pixels of noise when considering plane parallelism while the method not considering scene constraints has barely converged to the right solution in only 50% of the cases.

The results with varying sequence lengths (2 to 8 images) and 1.25 pixels of noise are reported in Fig. 3. For instance, Fig. 3-(a) shows that the quality of the 3D reconstruction remains in favor of using scene constraints even when 8 images are used. The same conclusion can be drawn with regards to the intrinsic parameters Fig. 3-(b) (our method) versus Fig. 3-(c) (Mendonça-Cipolla’s). The

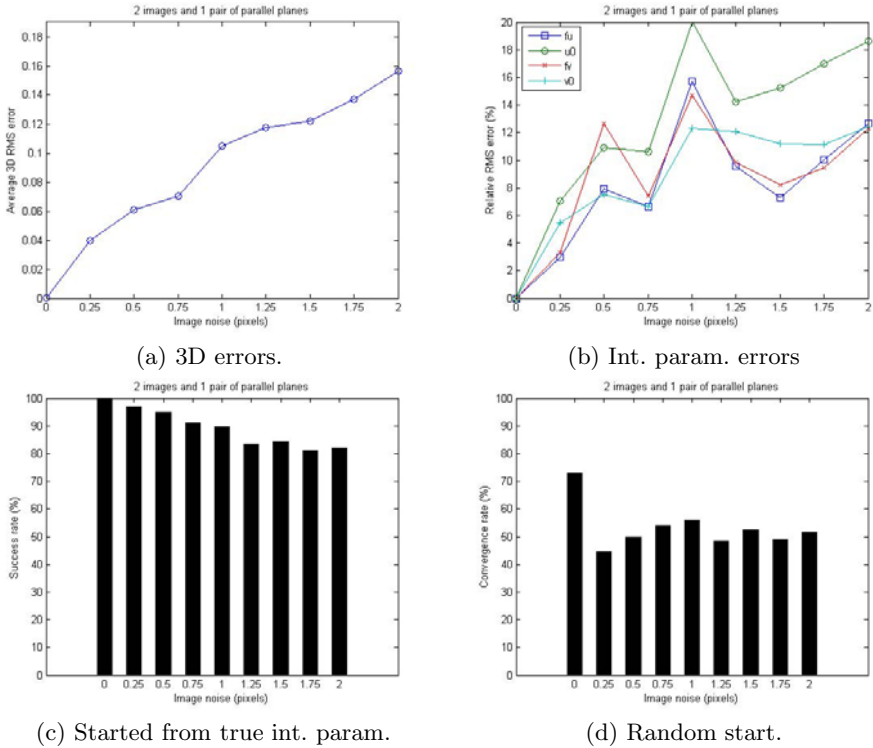
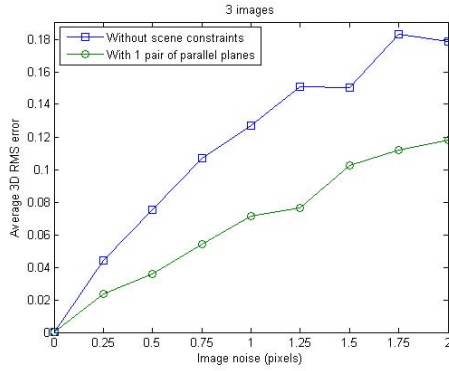


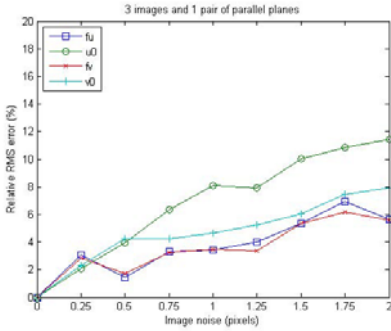
Fig. 1. Our method (1 pair of parallel planes): 2 images

convergence rates of the two methods are the same when using 6 or more images (Fig. 3-(d) and Fig. 3-(e)).

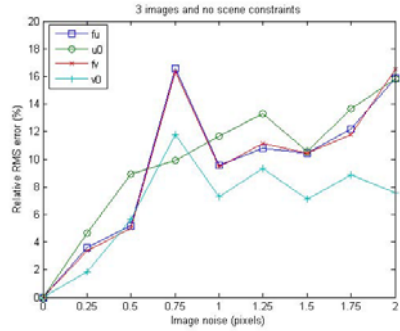
For comparison, the results obtained with the method presented here and those obtained with the stratified method [2] are reported in Fig. 4. Note that both methods use plane parallelism. The quality of the 3D reconstruction using (13) is better than the stratified method although the latter is based on a set of consistent projection matrices (usually leading to better results in scene-constraints-free methods). This may be due to the fact that the stratified method requires this extra-calculation of a consistent set of projective camera matrices in which, unlike when considering the images pairwise, errors or noise in one image might propagate to the entire set of camera matrices. Note that in these experiments, optimization of (13) has been randomly initialized and the convergence rates are given in Fig. 4-(b). For low levels of noise (typically < 0.75 pixel), the stratified method converges slightly more often than (13) to the right solution. However, for higher levels of noise (typically ≥ 0.75 pixel) our method has a better convergence rate. This can be explained by the fact that the stratified method finds it more difficult to single out the plane at infinity with the increasing noise. In addition, the step leading to an estimate of the intrinsic parameters from the plane at infinity becomes more sensitive to image noise.



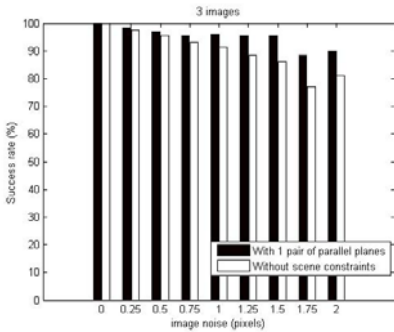
(a) 3D errors.



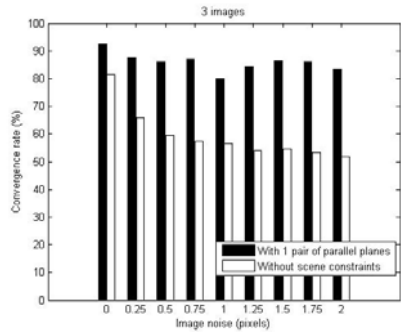
(b) Int. param. errors (our method).



(c) Int. param. errors (Mendonça-Cipolla).

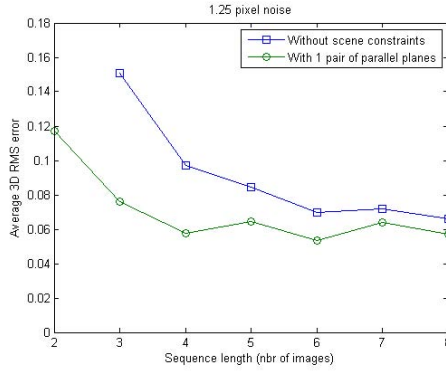


(d) Started from true int. param.

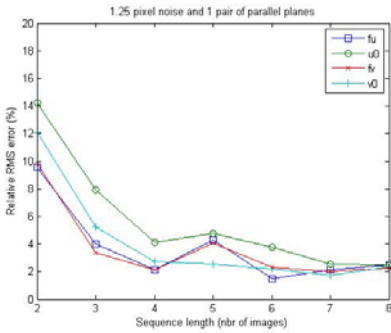


(e) Random start.

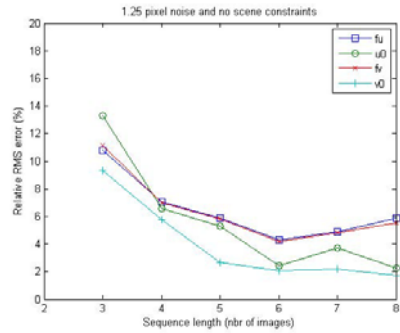
Fig. 2. Our method (1 pair of parallel planes) versus Mendonça-Cipolla's (without scene constraints): 3 images



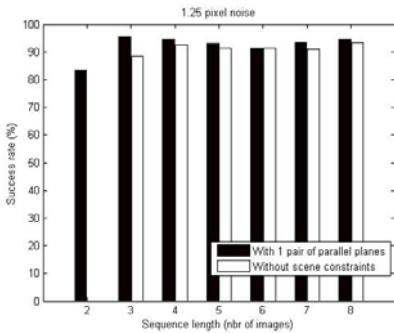
(a) 3D errors.



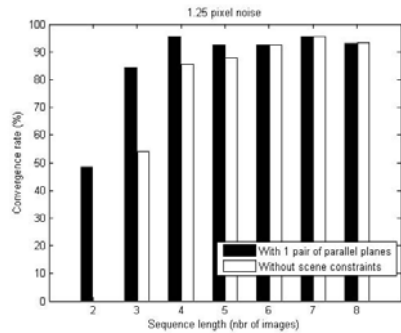
(b) Int. param. errors (our method).



(c) Int. param. errors (Mendonça-Cipolla).

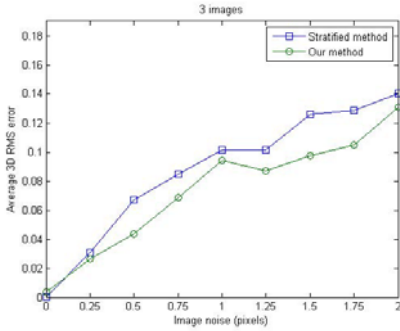


(d) Started from true int. param.

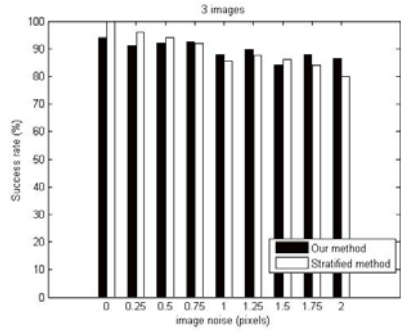


(e) Random start.

Fig. 3. Our method (1 pair of parallel planes) versus Mendonça-Cipolla's (without scene constraints): varying sequence lengths and 1.25 pixel noise



(a) 3D errors.



(b) Convergence rate.

Fig. 4. Our method (randomly started) vs. stratified method [2]: 3 images and 1 pair of parallel planes

5.2 Real Images

We have successfully tested our method on numerous real scenes of which we present here two examples: the "Patterns" scene, Fig. 5, and the "Desk" scene, Fig. 6. These images have been captured with a low-end Sony Cyber-shot DSC-S930. Only 2 images have been used in each example. The 48 points marked in Fig. 5 (middle) where matched across the two images of the "Patterns" scene

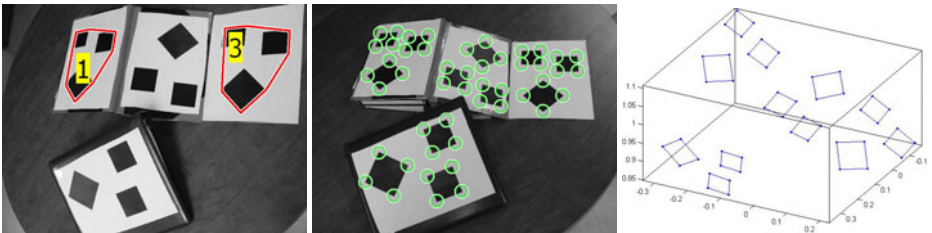


Fig. 5. The two images of the "Patterns" scene and its metric 3D reconstruction



Fig. 6. The two images of the "Desk" scene and its metric 3D reconstruction

while only the 28 points marked in Fig. 6 (middle) were matched across the images of the "Desk" scene. The planes labeled 1 and 3 in the left-hand side images in each figure were considered parallel. The recovered 3D metric structures of both scenes are given in Figs. 5 and 6 (right). The relative errors on the line segments ratios in both reconstructed scenes were found in the range 10%-18%.

6 Conclusion

We have proposed a new method that combines the self-calibration constraints on the essential matrix with plane parallelism constraints. The method allows the calibration of a camera from point correspondences as soon as two images are available (instead of three in the absence of scene constraints) and considers each pair of images independently. Our experiments have demonstrated that (1) the scene can be reconstructed using our method with a good accuracy even when using few images and high levels of image noise; (2) scene constraints contribute to significantly improve the quality of the 3D reconstruction and convergence in comparison with the works in [2] and [7].

References

1. Fusiello, A.: A New Autocalibration Algorithm: Experimental Evaluation. In: Skarbek, W. (ed.) CAIP 2001. LNCS, vol. 2124, pp. 717–724. Springer, Heidelberg (2001)
2. Habed, A., Amintabar, A., Boufama, B.: Affine camera calibration from homographies of parallel planes. In: Proceedings of the IEEE International Conference on Image Processing, pp. 4249–4252 (2010)
3. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision, 2nd edn. Cambridge University Press (2004) ISBN: 0521540518
4. Huang, T.S., Faugeras, O.D.: Some properties of the e matrix in two-view motion estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence 11(12), 1310–1312 (1989)
5. Huynh, D.Q., Heyden, A.: Scene point constraints in camera auto-calibration: An implementational perspective. Image and Vision Computing 23(8), 747–760 (2005)
6. Liebowitz, D., Zisserman, A.: Combining scene and auto-calibration constraints. In: Proceedings of the 7th International Conference on Computer Vision, Kerkyra, Greece, pp. 293–300 (1999)
7. Mendonça, P.R.S., Cipolla, R.: A simple technique for self-calibration. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado, USA, vol. 1, pp. 500–505 (June 1999)
8. Pollefeys, M., Van Gool, L.: Stratified self-calibration with the modulus constraint. IEEE Transactions on Pattern Analysis and Machine Intelligence 21(8), 707–724 (1999)

A Key Derivation Scheme for Hierarchical Access Control to JPEG 2000 Coded Images

Shoko Imaizumi¹, Masaaki Fujiyoshi², Hitoshi Kiya², Naokazu Aoki¹,
and Hiroyuki Kobayashi¹

¹ Graduate School of Advanced Integration Science, Chiba University,
1-33 Yayoicho, Inage-ku, Chiba-shi, Chiba, Japan

² Dept. of Information and Communication Systems, Tokyo Metropolitan University,
6-6 Asahigaoka, Hino-shi, Tokyo, Japan
imaizumi@chiba-u.jp, fujiyoshi-masaaki@tmu.ac.jp, kiya@sd.tmu.ac.jp,
{aoki,kobahiro}@faculty.chiba-u.jp
http://www.nd.chiba-u.jp/yugo-index_e.html

Abstract. This paper proposes a key derivation scheme to control access of JPEG 2000 (JP2) coded images, which consist of hierarchical scalability such as SNR, resolution levels, and so on. The proposed scheme simultaneously controls access to each level of scalability. The proposed scheme derives keys through hash chains, and each JP2 packet is enciphered with each individual key. By introducing combinations of a cyclic shift and a hash function, the proposed scheme manages only a single key for a JP2 image; whereas the conventional access control schemes having the above mentioned features manage multiple keys. The single managed key is not delivered to any user. The proposed scheme is also resilient to collusion attacks. Performance analysis shows the effectiveness of the proposed scheme.

Keywords: JPEG 2000, access control, key derivation, hash chain, cyclic shift.

1 Introduction

With the continuing growth in communication channels and terminals, scalable transmission, in which lower quality content is displayed by decompressing a certain portion of the codestream, is becoming popular. Scalable access control for the protection of scalable compressed images has been studied widely [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]. Security for JPEG 2000 (JP2) [11] is emphasized in JPEG 2000 Part8 [12], and JP2 coded images must be tightly secured.

A simple and straightforward way to realize hierarchical access control for JP2 coded images, consisting of several kinds of scalability, is the individual encipherment of each JP2 packet. This approach, however, must manage a large number of keys, given the large number of JP2 packets in a JP2 coded image.

Scalable access control schemes have also been proposed for JP2 coded images [3, 4, 5, 6, 7, 8, 9, 10]. These schemes use one- or multi-dimensionally hierarchical scalability provided by coding technologies, so that the user can obtain

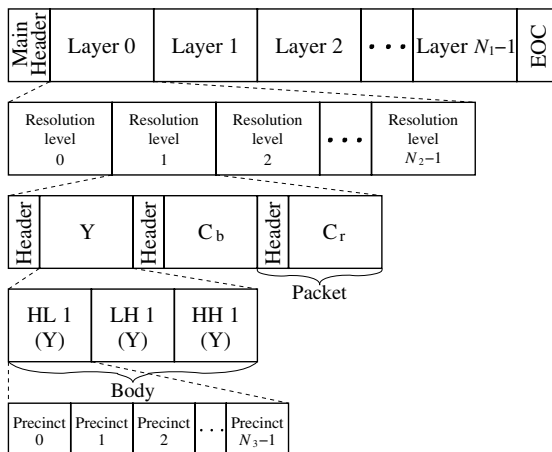


Fig. 1. JP2 codestream with color components, Y, C_b , and C_r . The progression order is LRCP.

an image or a video at the permitted quality from one common codestream. The hash chain [13] has also been introduced to several schemes for reduction of the number of managed keys and the keys delivered to the user (delivered keys) [6, 7, 8, 9, 10]. Although these hash chain-based access control schemes are effective for reduction of the number of keys, the number of managed keys increases, depending not only on the kinds of scalability, but also on the depth of the hierarchy in each scalability.

This paper proposes an efficient key derivation scheme for hierarchical access control to JP2 coded images in which several kinds of scalability exist. By introducing combinations of a cyclic shift and a hash function, the number of managed keys is reduced to one. The managed key is not delivered to any user, providing security against key leakage. The proposed scheme is also resilient to collusion attacks, in which malicious users illegally access an image at higher quality than that allowed by their access rights.

2 JP2 Codestream and Hierarchical Access Control

This section briefly describes JP2 codestream structure [11] and scalable access control for JP2 coded images. It also summarizes the requirements for hierarchical access control methods by introducing four conventional methods [7, 8, 9, 10] to clarify the aim of this work.

2.1 JP2 Codestream

Fig. 1 outlines a JP2 codestream using YC_bC_r as the color space. JP2 supports five different progression orders that are orders of scalability, and the default order, that is also used in Fig. 1, is LRCP (Layer-Resolution-Component-Precinct). It is primarily progressive by quality.

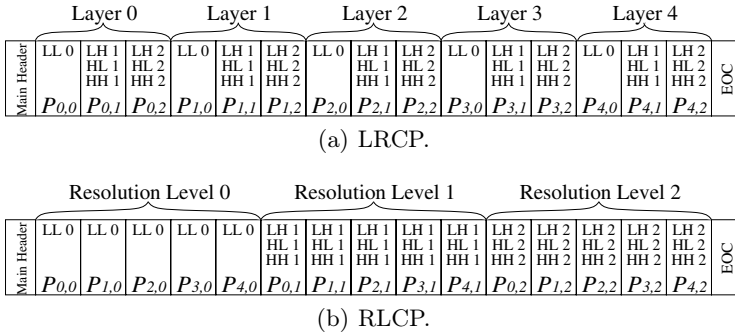


Fig. 2. Progression orders in a grayscale image with five layers and three resolution levels, i.e., $N_1 = 5$ and $N_2 = 3$

Layers are in order of SNR in which each layer is composed of data for resolution levels. If the original image has color components, each resolution level has Y , C_b , and C_r components. Resolution level zero only contains the LL data, whereas the other levels contain three subbands; HL, LH, and HH. These subbands have precincts that have non-hierarchically positional information. Thus, a color JP2 codestream has three kinds of hierarchical scalability; layer, resolution level, and components, whereas a grayscale JP2 codestream has two; layer and resolution level. Each JP2 packet is composed of a header and a body and contains partial data for each subband.

Fig. 2 lists examples of JP2 codestreams with LRCP and RLCP progression orders. Both have five layers and three resolution levels, which are represented as $N_1 = 5$ and $N_2 = 3$, respectively, in this paper. Hereafter, P_{n_1, n_2} is the JP2 packet at the n_1 -th layer and n_2 -th resolution level.

2.2 Hierarchically Access Control

Fig. 3 outlines an example of scalable decoding in which different image products are obtained by decompression in many ways, where $N_1 = 5$ and $N_2 = 3$. In this example, the decoded image is grayscale. It is noted that this representation holds regardless of progression orders. The original image is compressed at quality $Q_{4,2}$, and the image at $Q_{4,2}$ is obtained by decompressing all packets. To produce the image at $Q_{1,1}$, four packets $P_{0,0}$, $P_{0,1}$, $P_{1,0}$, and $P_{1,1}$ are decompressed. Thus, a scalable access control method for JP2 should encipher a JP2 codestream packet-by-packet using $N_1 \times N_2$ different keys. Access control for JP2 coded images should encipher the packet body but does not encipher the packet headers.

2.3 Requirements

This section describes two requirements for hierarchical access control for JP2 coded images, i.e., collusion attack-resilience and the less number of managed keys.

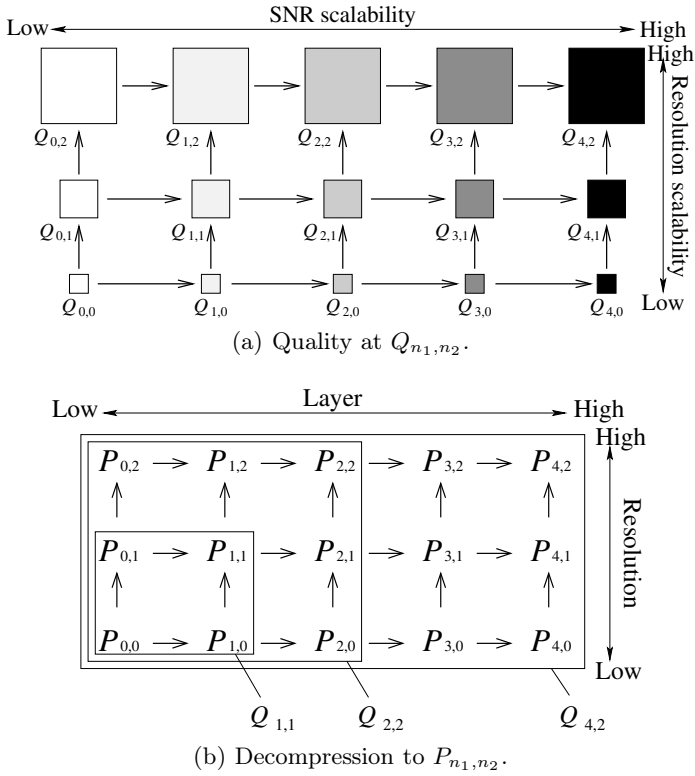


Fig. 3. Hierarchical decomposition of a grayscale image ($N_1 = 5$ and $N_2 = 3$)

Collusion Attack-Resilience. A collusion attack is made by multiple users to obtain an image with higher quality than that allowed by their access rights, and the conventional scheme [7] allows users to collude. In Fig. 4, the arrows indicate direction of key derivation. The key for packet P_{n_1, n_2} is K_{n_1, n_2} , and $K_{4,2}$ is the managed key. As shown in Fig. 5, the managed key $K_{4,2}$ is divided into two partial keys $K_{PK_1(4)}$ and $K_{PK_2(2)}$. Each partial key is allocated to each hierarchy, and the partial keys $K_{PK_1(n_1)}$ and $K_{PK_2(n_2)}$ for key K_{n_1, n_2} are derived from previous partial keys $K_{PK_1(n_1+1)}$ and $K_{PK_2(n_2+1)}$, using hash chains [13]. By concatenating them, $K_{n_1, n_2} = (K_{PK_1(n_1)} \parallel K_{PK_2(n_2)})$, is derived.

In Fig. 4 (a), Alice is allowed to access the image at $Q_{0,2}$ and receives key $K_{0,2}$, which is consisting of two partial keys $K_{PK_1(0)}$ and $K_{PK_2(2)}$. She can derived keys $K_{0,1}$ and $K_{0,0}$ and decipher $P_{0,2}$, $P_{0,1}$, and $P_{0,0}$. Whereas, Bob, in Fig. 4 (b), receives $K_{4,0}$, which is consisting of $K_{PK_1(4)}$ and $K_{PK_2(0)}$, and derives $K_{3,0}$, $K_{2,0}$, $K_{1,0}$, and $K_{0,0}$ to decipher $P_{4,0}$, $P_{3,0}$, $P_{2,0}$, $P_{1,0}$, and $P_{0,0}$ for access the image at $Q_{4,0}$. In this scheme, they are possible to illegally derive $K_{4,2}$ by using $K_{PK_1(4)}$ and $K_{PK_2(2)}$, so they can decipher all packets as shown in Fig. 4 (c) and access the image at $Q_{4,2}$. The proposed scheme is resistant to collusion attacks.

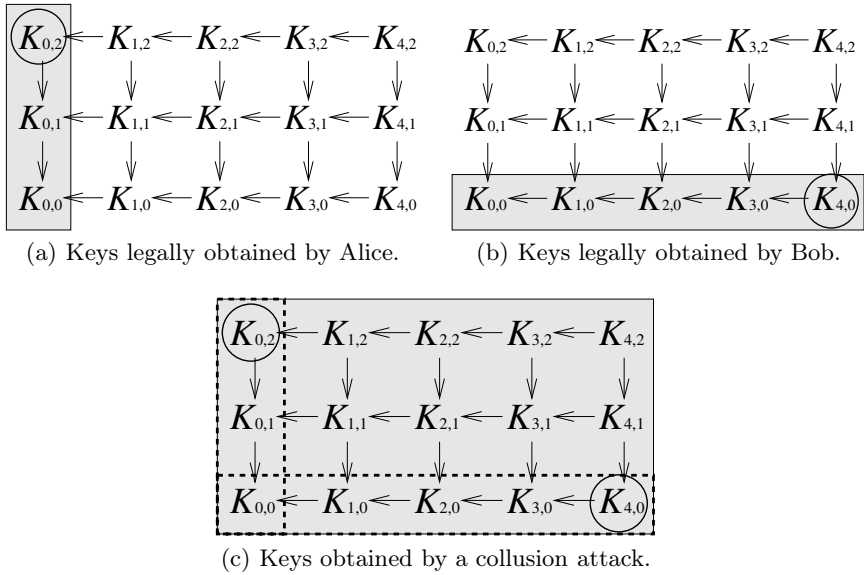


Fig. 4. Alice and Bob’s collusion attack in the vulnerable scheme [7] (the shaded are obtained)

The Less Number of Managed Keys. Although a hierarchical access control method requires $N_1 \times N_2$ of keys as mentioned in Sect. 2.2, three schemes that manage less keys and subordinately derive $N_1 \times N_2$ of keys from the managed keys have been proposed [8, 9, 10].

The first scheme [8] controls access to JP2 codestreams according to the hierarchy in the prior scalability. This scheme, Scheme I hereafter, subordinately derives keys from the managed key using hash chains [13]. It, thus, requires five managed keys and five codestreams for five progression orders. The number of managed keys in Scheme I, $N_{m,I}$, is

$$N_{m,I} = 5. \tag{1}$$

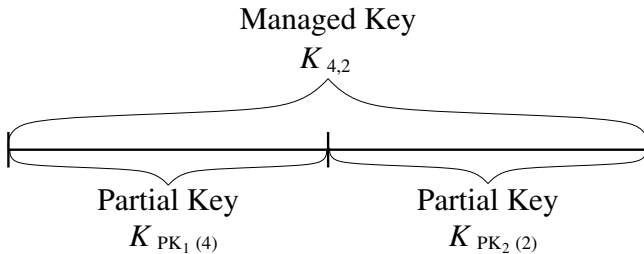


Fig. 5. Managed key consisting of two partial keys [7]

The second and third schemes [9], Scheme II and Scheme III hereafter, simultaneously control access in every hierarchical scalability with a single codestream. The number of managed keys in Scheme II, $N_{m,II}$, is

$$N_{m,II} = \min(N_1, N_2), \tag{2}$$

and the number of managed keys in Scheme III, $N_{m,III}$, is

$$N_{m,III} = N_1 + N_2 - 1, \tag{3}$$

whereas the proposed scheme needs only a single managed key.

3 Proposed Scheme

This section proposes a new scheme for access control to JP2 coded images that reduces the number of managed keys to one. The proposed scheme simultaneously controls access in every hierarchical scalability with a single managed key and a single managed codestream. The proposed scheme is resistant to collusion attacks as Schemes I, II, and III.

3.1 Key Derivation and Encipherment of Codestream

As an example of JP2 codestreams for explanation, the proposed scheme assumes the JP2 codestream shown in Fig. 2, where it is composed of five layers ($N_1 = 5$) and three resolution levels ($N_2 = 3$). The proposed scheme controls access regardless of progression orders.

Fig. 6 shows a new key derivation order, where K_{n_1,n_2} is the key for packet P_{n_1,n_2} . This order is resilient to collusion attacks. It is noted that key K_m is the single managed key.

Firstly in the proposed scheme, key $K_{4,2}$ is derived from K_m as

$$K_{4,2} = h(s(K_m)), \tag{4}$$

where $s(\cdot)$ is a cyclic shift and $h(\cdot)$ is a cryptographic one-way hash function. Replacing the combination of $s(\cdot)$ and $h(\cdot)$ with $f(\cdot)$, Eq. (4) is represented as

$$K_{4,2} = f(K_m). \tag{5}$$

Similarly, keys $K_{4,1}$ and $K_{4,0}$ are derived by

$$K_{4,n_2} = f^{3-n_2}(K_m), \quad n_2 = 1, 0, \tag{6}$$

respectively, where $f^\alpha(\beta)$ represents that $f(\cdot)$ is applied to β recursively α times. The combinations of a cyclic shift and a hash function $f(\cdot)$ are shown with dashed arrows in Fig. 6.

Meanwhile, keys $K_{n_1,2}$ ($n_1 = 3, 2, 1, 0$) are derived by a hash chain. In this example, these keys are given as

$$K_{n_1,2} = h^{4-n_1}(K_{4,2}), \quad n_1 = 3, 2, 1, 0, \tag{7}$$

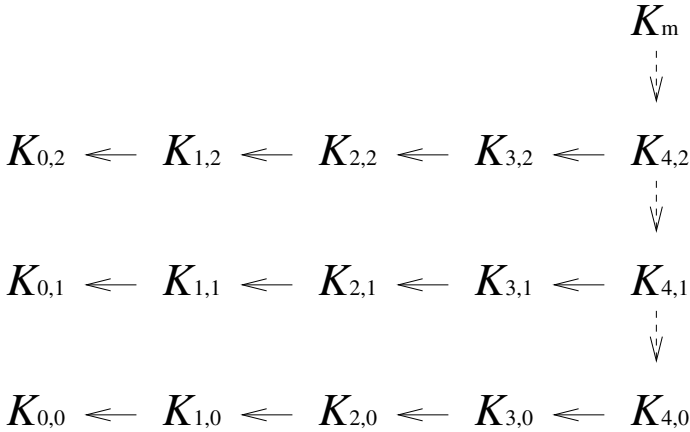


Fig. 6. Key derivation to control access to the JP2 codestream having five layers and three resolution levels ($N_1 = 5$ and $N_2 = 3$). K_{n_1, n_2} is the key for JP2 packet P_{n_1, n_2} . A solid arrow is an hash function and a dashed arrow is a combination of a cyclic shift and a hash function.

respectively, where $h^\alpha(\beta)$ represents that cryptographic one-way hash function $h(\cdot)$ is applied to β recursively α times. Similarly, keys $K_{n_1, 1}$ and $K_{n_1, 0}$ are derived by

$$\begin{aligned}
 K_{n_1, n_2} &= h^{4-n_1}(K_{4, n_2}), \\
 n_1 &= 3, 2, 1, 0, \quad n_2 = 1, 0,
 \end{aligned}
 \tag{8}$$

respectively. The hash chains are shown with solid arrows in Fig. 6.

By introducing a combination of a cyclic shift and a hash function shown in Eq. (4), all keys K_{n_1, n_2} for JP2 packets P_{n_1, n_2} are derived from single managed key K_m .

With key K_{n_1, n_2} , JP2 packet P_{n_1, n_2} in the JP2 codestream is enciphered, where $n_1 = 0, 1, \dots, N_1 - 1$ and $n_2 = 0, 1, \dots, N_2 - 1$. It is noted that any arbitrary symmetric encipher algorithm can be used in the proposed scheme.

3.2 Decipherment and Decompression of Codestream

Here, it is considered that a user is allowed to access the image with quality $Q_{2,2}$, c.f. Fig. 3. The user receives keys $K_{2,2}$, $K_{2,1}$, and $K_{2,0}$ as shown in Fig. 7(a). To decompress the image at $Q_{2,2}$, the user needs to decipher nine packets $P_{0,0}$, $P_{0,1}$, $P_{0,2}$, $P_{1,0}$, $P_{1,1}$, $P_{1,2}$, $P_{2,0}$, $P_{2,1}$, and $P_{2,2}$. The six keys $K_{0,0}$, $K_{0,1}$, $K_{0,2}$, $K_{1,0}$, $K_{1,1}$, and $K_{1,2}$ that the user needs are derived from the delivered keys $K_{2,2}$, $K_{2,1}$, and $K_{2,0}$ as

$$\begin{aligned}
 K_{n_1, n_2} &= h^{2-n_1}(K_{2, n_2}), \\
 n_1 &= 1, 0, \quad n_2 = 2, 1, 0.
 \end{aligned}
 \tag{9}$$

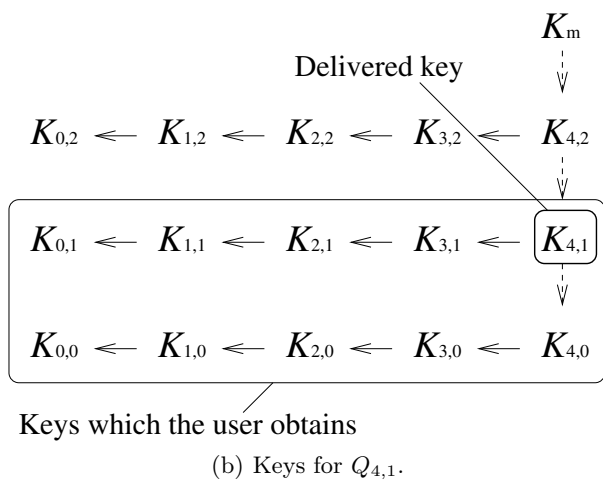
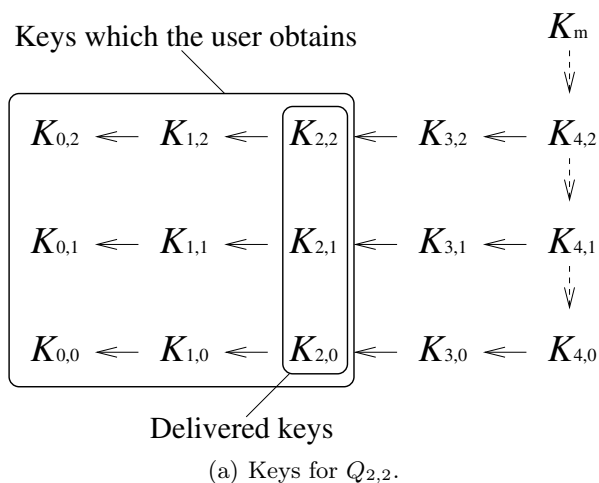


Fig. 7. Delivered and derived keys that the user needs to decompress the image at certain quality

By using keys $K_{0,0}$, $K_{0,1}$, $K_{0,2}$, $K_{1,0}$, $K_{1,1}$, $K_{1,2}$, $K_{2,0}$, $K_{2,1}$, and $K_{2,2}$, corresponding packets are deciphered and decompressed to present the image at $Q_{2,2}$.

As another example, it is assumed that a user can access the image with quality $Q_{4,1}$. The user receives single key $K_{4,1}$ as shown in Fig. 7(b). To access the image at $Q_{4,1}$, the user has to obtain ten keys $K_{0,0}$, $K_{0,1}$, $K_{1,0}$, $K_{1,1}$, $K_{2,0}$, $K_{2,1}$, $K_{3,0}$, $K_{3,1}$, $K_{4,0}$, and $K_{4,1}$. First, $K_{4,0}$ is derived from the delivered key $K_{4,1}$ as

Table 1. Comparisons in terms of the number of managed keys and delivery of managed keys

	Proposed Scheme I [8]	Scheme II [9]	Scheme III [10]	
The number of managed keys	1	5	$\min(N_1, N_2)$	$N_1 + N_2 - 1$
Delivery of managed keys	No	Yes	Yes	Yes

$$\begin{aligned}
 K_{4,0} &= h(s(K_{4,1})) \\
 &= f(K_{4,1}),
 \end{aligned}
 \tag{10}$$

which is the combination of a cyclic shift and a hash function. Then, eight keys K_{n_1, n_2} ($n_1 = 0, 1, 2, 3, \quad n_2 = 1, 0$) are derived by

$$\begin{aligned}
 K_{n_1, n_2} &= h^{4-n_1}(K_{4, n_2}), \\
 n_1 &= 3, 2, 1, 0, \quad n_2 = 1, 0,
 \end{aligned}
 \tag{11}$$

and the user can obtain the ten keys for the ten packet.

3.3 Features

This section verifies that the proposed scheme meets requirements described in Sect. 2.3. The proposed scheme is evaluated by comparing with the conventional schemes [7, 8, 9, 10] which use hash chains [13] only.

Collusion Attack-Resistance. The proposed scheme is resilient to collusion attacks as well as the conventional schemes [8, 9, 10], i.e., Schemes I, II, and III, while the conventional scheme [7] is naive for collusion attacks.

Alice and Bob appeared in Sect. 2.3 reappear here. Since Alice can access the image at $Q_{0,2}$, she receives keys $K_{0,2}$, $K_{0,1}$, and $K_{0,0}$. Bob receives single key $K_{4,0}$ to access the image at $Q_{4,0}$. Bob derives $K_{3,0}$, $K_{2,0}$, $K_{1,0}$, and $K_{0,0}$ from his delivered key $K_{4,0}$ by using Eq. (8). They obtain seven valid keys $K_{0,0}$, $K_{0,1}$, $K_{0,2}$, $K_{1,0}$, $K_{2,0}$, $K_{3,0}$, and $K_{4,0}$, but they can not derives any keys which they are not permitted to derive from these seven keys.

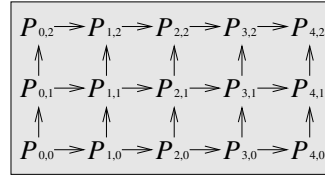
Thus, the proposed scheme is enough resistant to collusion attacks, though this paper does not explicate all pattern of collusion attacks.

Managed keys. Table 1 shows the results of comparisons in terms of the number of managed keys and delivery of managed keys. The proposed scheme manages only a single key regardless of the kinds of scalability and the depth of the hierarchy in each scalability, whilst Scheme I [8] must manage five keys and Scheme II [9] must manage keys as many as the minimum depth of hierarchy of two scalabilities. The number of managed keys in Scheme III [10] is just about the sum of the depth of two hierarchical scalabilities.

The single managed key is not delivered to any user in the proposed scheme, whereas the managed keys are delivered to some users in Schemes I, II, and III.



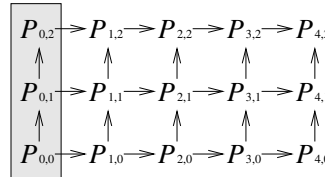
(a) Fully decompressed ($Q_{4,2}$). PSNR: 36.68 dB.



(b) Decoded packets for (a).



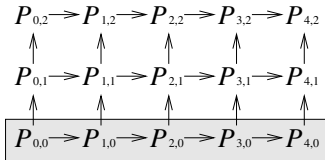
(c) Alice's ($Q_{0,2}$). PSNR: 27.71 dB.



(d) Decoded packets for (c).



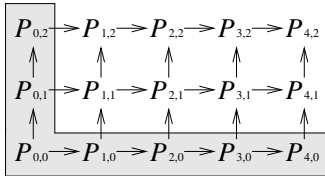
(e) Bob's ($Q_{4,0}$). PSNR: 29.51 dB.¹



(f) Decoded packets for (e).



(g) Colluded. PSNR: 30.18 dB.



(h) Decoded packets for (g).

Fig. 8. Image examples. 512×512 -sized lena is compressed. Five 0.1 bits/pixel-rate layers ($N_1 = 5$) and three resolution levels ($N_2 = 3$).

¹ Decompression of the LL subband and other subbands filled with zero.

4 Experimental Results

Grayscale image “lena” is compressed by Kakadu to generate a codestream with five layers ($N_1 = 5$) and three resolution levels ($N_2 = 3$). The bitrate of a layer is 0.1 bits/pixel, and Fig. 8 (a) shows the fully decompressed image, i.e., at quality $Q_{4,2}$. Alice can access the image with quality $Q_{0,2}$ shown in Fig. 8 (c), and Bob obtains the image shown in Fig. 8 (e) as $Q_{4,0}$. In the proposed scheme, Alice and Bob illegally derive the image shown in Fig. 8 (g). Since no illegally deciphered packet contributes the quality of this image, two users do not benefit from the collusion attack. Simulations with other images give similar results.

5 Conclusion

This paper has proposed a new key derivation scheme for access control to JP2 coded images in which combinations of a cyclic shift and a hash function are employed. The proposed scheme manages a single key and the single managed key is not delivered to any user. The proposed scheme also prevents malicious users to collude for accessing an images at higher quality much than that allowed by their permission.

References

1. Xie, D., Kuo, C.C.J.: Multimedia data encryption via random rotation in partitioned bit streams. In: Proc. IEEE ISCAS, pp. 5533–5536 (2005)
2. Zhang, Z., Sun, Q., Wong, W.C., Apostolopoulos, J., Wee, S.: Rate-distortion-authentication optimized streaming of authenticated video. *IEEE Trans. Circuits Syst. for Video Technol.* 17, 544–557 (2007)
3. Grosbois, R., Gerbelot, P., Ebrahimi, T.: Authentication and access control in the JPEG 2000 compressed domain. In: Proc. SPIE, vol. 4472, pp. 95–104 (2001)
4. Haggag, A., Ghoneim, M., Lu, J., Yahagi, T.: Progressive encryption and controlled access scheme for JPEG 2000 encoded images. In: Proc. IEEE ISPACS, pp. 895–898 (2006)
5. Shahid, Z., Chaumont, M., Puech, W.: Selective and scalable encryption of enhancement layers for dyadic scalable H.264/AVC by scrambling of scan patterns. In: Proc. IEEE ICIP, pp. 1273–1276 (2009)
6. Won, Y.G., Bae, T.M., Ro, Y.M.: Scalable Protection and Access Control in Full Scalable Video Coding. In: Shi, Y.Q., Jeon, B. (eds.) IWDW 2006. LNCS, vol. 4283, pp. 407–421. Springer, Heidelberg (2006)
7. Joye, M., Yen, S.M.: One-Way Cross-Trees and Their Applications. In: Naccache, D., Paillier, P. (eds.) PKC 2002. LNCS, vol. 2274, pp. 346–356. Springer, Heidelberg (2002)
8. Wu, Y., Ma, D., Deng, R.H.: Progressive protection of JPEG 2000 codestreams. In: Proc. IEEE ICIP, pp. 3447–3450 (2004)
9. Imaizumi, S., Fujiyoshi, M., Kiya, H.: Efficient collusion attack-free access control for multidimensionally hierarchical scalability content. In: Proc. IEEE ISCAS, pp. 505–508 (2009)

10. Imaizumi, S., Fujiyoshi, M., Abe, Y., Kiya, H.: Collusion attack-resilient hierarchical encryption of JPEG 2000 codestreams with scalable access control. In: Proc. IEEE ICIP, pp. II-137–II-140 (2007)
11. Information technology — JPEG 2000 image coding system – Part 1: Core coding system. ISO/IEC IS-15444-1 (2004)
12. Information technology — JPEG 2000 image coding system – Part 8: Secure JPEG 2000. ISO/IEC IS-15444-8 (2007)
13. Lamport, L.: Password authentication with insecure communication. *Communications of the ACM* 24(11), 770–772 (1981)

Bifocal Matching Using Multiple Geometrical Solutions

Miguel Carrasco¹ and Domingo Mery²

¹ Escuela de Informática y Telecomunicaciones
Universidad Diego Portales
Ejército 441, Santiago de Chile
`miguel.carrasco@mail.udp.cl`

² Departamento de Ciencia de la Computación
Pontificia Universidad Católica de Chile
Av. Vicuña Mackenna 4860(143), Santiago de Chile
`dmery@ing.puc.cl`

Abstract. Determining point-to-point correspondence in multiple images is a complex problem because of the multiple geometric and photometric transformations and/or occlusions that the same point can undergo in corresponding images. This paper presents a method of point-to-point correspondence analysis based on the combination of two techniques: (1) correspondence analysis through similarity of invariant features, and (2) combination of multiple partial solutions through bifocal geometry. This method is quite novel because it allows the determination of point-to-point geometric correspondence by means of the intersection of multiple partial solutions that are weighted through the MLESAC algorithm. The main advantage of our method is the extension of the algorithms based on the correspondence of invariant descriptors, generalizing the problem of correspondence to a geometric model in multiple views. In the sequences used we got an F-score = 97% at a distance of less than 1 pixel. These results show the effectiveness of the method and potentially can be used in a wide range of applications.

Keywords: computer vision, multiple view geometry, correspondence problem, tracking.

1 Introduction

The point-to-point correspondence analysis between two images made of the same scene is a very relevant problem in the computer vision community. Problems such as 3D reconstruction, robotic navigation, tracking in multiple views, estimation of transformation matrices and homographies, among others, are some of the applications that require the precise determination of correspondences. Correspondence analysis consists basically in determining a set of points in an image such that they are identified as the same in other images of the same scene. This situation is described clearly in Fig.1, which shows a possible corresponding point in three images of the same object. Different approaches for

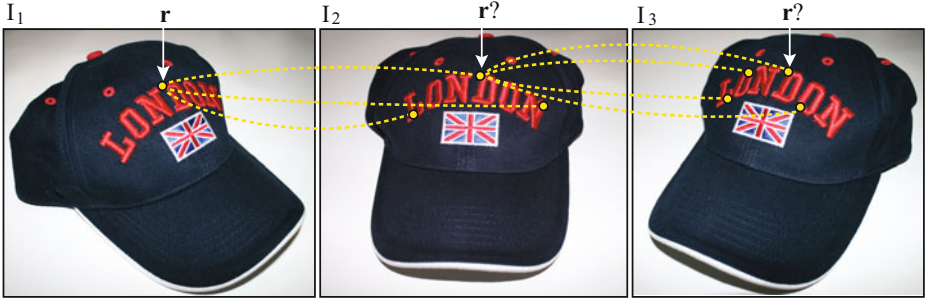


Fig. 1. General problem of correspondences in multiple views. Main objective: determining the correspondence of point r of image I_1 in corresponding images I_2 e I_3 .

solving the correspondence matching have been developed. Some of them are, for example, methods based on the analysis of invariant descriptors [3,11], estimation of affine transformations, homographies and estimation of perspective transformations [4,8], epipolar geometry analysis [14,19], and methods based on optical flow [1]. In general, all these methods differ in the type of motion of the objects contained in a video sequence.

In spite of the large number of methods designed to solve this problem, the correspondence of images with very wide viewing angles has not been solved completely. This problem is commonly found when two or more independent cameras located at different positions and with wide perspective angles are used [5]. It is common to use techniques based on the analysis of invariant descriptors. Thanks to the invariance it is possible to solve and generalize the point-to-point correspondence problem by providing an extension of the methods based on stereo vision. That correspondence takes place as a function of the points of interest detected previously by some algorithm for detecting regions of interest [12]. However, when the point of interest does not correspond to a point detected by the current saliency techniques [10,18], how can we determine its corresponding pair in the other images? In this case, the previous methods do not ensure finding a correct correspondence because they are designed to maximize their performance only in the regions of interest detected by the method, and not necessarily in other regions. To avoid the problems of the correspondence methods mentioned previously, in this research we propose a new method to determine the point-to-point correspondence, particularly when the displacement angles are wide. Furthermore, since we use a geometric model that is independent of the objects, it is possible to determine the position of corresponding points in those views in which the point may be occluded. Graphically, we propose to solve the problem of Fig. 1. Given a point r in the image I_1 , the objective is to determine a corresponding point in the image I_2 .

In the following sections we detail our methodology for estimating the correspondence matrices in multiple views in uncalibrated sequences. The rest of the document is organized in the following sections: section 2 includes a description

of the proposed method; section 3 includes the experiments and results; and finally, section 4 presents the conclusions and future work.

2 Proposed Methodology

In general, all the search methods for the fundamental matrix [9] have the purpose of finding the best model generated from a random set of pairs in correspondence through an error minimization process [7,20,16]. This process can take place, for example, by means of a sampling consensus known as RANSAC [7], or the likelihood maximization in MLESAC by random sampling [16]. Both methods, as well as the improvements proposed by Torrdooff and Murray [15], have been shown to be efficient methods for finding the fundamental matrices and perspectives in problems of computer vision. In the case of two views, the objective of the minimization process is to determine an epipolar single line in order to find an optimum epipole [14]. However, these methods have been designed for problems in which there is a considerable number of erroneous correspondences, so the random search for hypotheses has the objective of determining the quality of each selected hypothesis and in that way reduce the selection of erroneous correspondences [16]. But what happens when there is a large number of correctly estimated correspondences? Is the best hypothesis the only solution that can be used? To answer these questions, below we present a new method for determining the point-to-point correspondence in two views in a geometric way.

2.1 Correspondence in Two Views

One of the most widely studied problems in computer vision is the geometric relation that exists between two corresponding images. A first step to solve this problem is to determine a set of point-to-point correspondences that estimate the geometric relations present in both images. In general, the problem of analysis in two views consists of how to determine the geometric relations of a 3D point and its projections on 2D planes. In what follows we will introduce the notation that relates the points in both images and the geometry that defines them. First, let \mathbf{P} be a point in 3D space. In our example, point \mathbf{P} is located in the upper corner of the 3D cube of Fig. 2. Second, let \mathbf{C}_1 and \mathbf{C}_2 be the optical centers of two cameras. For the following analysis, assume that we capture an image from the optical center \mathbf{C}_1 , which generates image \mathbf{I}_1 . Also, if we capture an image from the optical center \mathbf{C}_2 , we generate image \mathbf{I}_2 . According to this configuration, if we project a ray from center \mathbf{C}_1 to point \mathbf{P} , point \mathbf{r} is generated on the 2D plane of image \mathbf{I}_2 . Similarly, if we project a ray from center \mathbf{C}_2 to point \mathbf{P} , point \mathbf{m} is generated, defined on the 2D plane of image \mathbf{I}_2 . This relation implies that both rays intersect at a single point \mathbf{P} defined in 3D space and its projections are on the \mathbf{I}_1 and \mathbf{I}_2 planes. In this way, points \mathbf{r} and \mathbf{m} correspond to a projection of point \mathbf{P} . In the ideal case both points are corresponding, since they were generated from a single point, in this case point \mathbf{P} . On the contrary, if we do not know the existence of point \mathbf{P} we cannot assure that the correspondence is true.

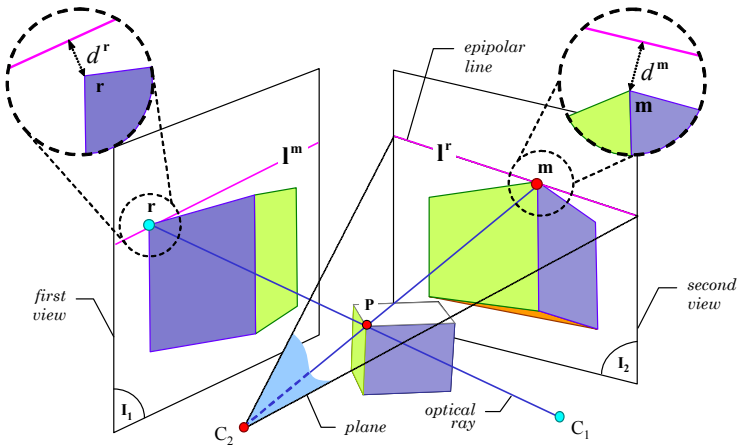


Fig. 2. General epipolar geometry of a 3D object and its projections

The latter situation is what normally occurs in point-to-point correspondence problems. In what follows we will denote by $\{\mathbf{r} \leftrightarrow \mathbf{m}\}$, when points \mathbf{r} and \mathbf{m} are corresponding, and as $\{\mathbf{r} \mapsto \mathbf{m}\}$ when the relation is hypothetical, i.e., we do not know if the relation is true or false and we want to find out.

A conventional way of proving the relation between points \mathbf{r} and \mathbf{m} is through the *fundamental matrix* \mathbf{F} [9,14]. Formally, the fundamental matrix encapsulates the intrinsic geometry of two views, called *epipolar geometry*. For its determination it is necessary to know a minimum set of correspondences in both views. The main relation that establishes it is: given a pair of $\{\mathbf{r} \leftrightarrow \mathbf{m}\}$ correspondences, they always satisfy the following epipolar restriction:

$$\mathbf{m}^\top \cdot \mathbf{F} \cdot \mathbf{r} = 0,$$

Unfortunately, this relation is valid for all the points that are found at the intersection of the projection plane of the optical center \mathbf{C}_2 and the \mathbf{I}_2 plane. The line generated by that intersection is known as *epipolar line* [9]. Since point \mathbf{r} belongs to the plane of the optical center \mathbf{C}_2 , we say that the epipolar line in the second view \mathbf{I}_2 is correspondent with point \mathbf{r} in the first view \mathbf{I}_1 . According to this analysis, it is not possible to determine a biunivocal relation between points \mathbf{r} and \mathbf{m} using only an epipolar line. Various methods for estimating the fundamental matrix have been developed in recent years, e.g. [9,2]. Regardless of the method for estimating the fundamental matrix, once it is determined it is possible to calculate the epipolar line described above, as shown in Fig. 2. Formally, let \mathbf{I}^r be the epipolar line of point \mathbf{r} located in view \mathbf{I}_2 , defined as $\mathbf{I}^r = \mathbf{F} \cdot \mathbf{r}$. Let us assume that points \mathbf{r} and \mathbf{m} are corresponding. Therefore, they must be on epipolar lines \mathbf{I}^r and \mathbf{I}^m because they belong to the same plane. That is, $\mathbf{I}^m \cdot \mathbf{r} = 0$ and $\mathbf{I}^r \cdot \mathbf{m} = 0$. However, in practice the measurements of both views are not precise, and this implies that the epipolar lines do not necessarily

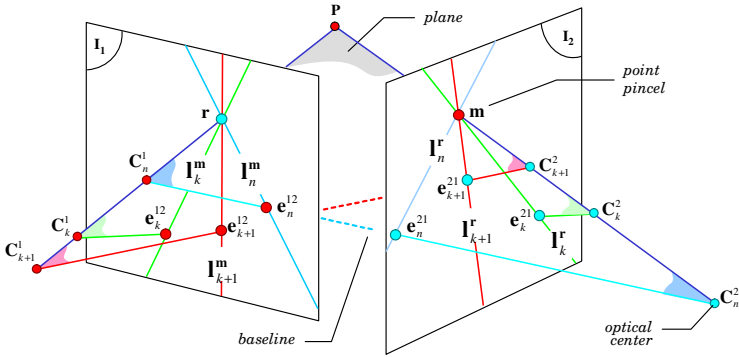


Fig. 3. Family of n multiple epipolar lines in two views from multiple epipoles

intersect the points in correspondence. This error is reflected in the Euclidian distances d^r and d^m between the real point and the epipolar line, in which $d^r > 0$ and $d^m > 0$ (Fig. 2). Both distances should be minimal so that the projections are correct. To minimize that error it is necessary to determine a set of correspondences that minimize a reprojection error. Normally the correspondence selection methods use some distance measure or probabilistic value to carry out that minimization. In some cases the error is generated by optical distortions belonging to the lenses or by Gaussian noise present in the acquisition of the coordinates in correspondence.

An important point related to the estimation of the fundamental matrix is its dependence with respect to the set of correspondences used; i.e., for every set of correspondences a new fundamental matrix is determined. Even when the fundamental matrices are different, all of them remain valid provided $|\mathbf{F}| = 0$. However, every fundamental matrix has associated with it a level of error due to the inaccuracies of the set of correspondences used. In spite of this error, the use of multiple fundamental matrices has two important advantages. (1) Every new fundamental matrix defines a new epipole position in the \mathbf{I}_1 and \mathbf{I}_2 planes. (2) The intersection of the epipole and the hypothetical point in correspondence (\mathbf{r} or \mathbf{m}) generates a new epipolar line. Taking into account the two previous properties, let us assume that we choose k sets in correspondence, where $k \in [1, \dots, n]$ and n is the maximum number of sets in correspondence. According to Fig. 3 the \mathbf{e}_k^{12} and \mathbf{e}_k^{21} epipoles are defined as the points of intersection between the baseline of the optical centers \mathbf{C}_k^1 and \mathbf{C}_k^2 , and the \mathbf{I}_1 and \mathbf{I}_2 planes, respectively.

In this case, for the model proposed in Fig. 3 we assume that the position of point \mathbf{P} is fixed. To illustrate the process in two views, we will assume that given a point \mathbf{r} in the first view, there is a corresponding point in the second view. Since we do not know that correspondence, in our example we will assume that there are three hypothetical corresponding points, that we will call \mathbf{m} , \mathbf{n} , and \mathbf{p} . As shown in Fig. 4, for the first set of correspondences the epipolar line \mathbf{I}_1^r intersects points \mathbf{m} , \mathbf{n} , and \mathbf{p} in the second view \mathbf{I}_2 . Therefore, let Θ be the set of hypothetical correspondences, where $\Theta = \{\{\mathbf{r} \mapsto \mathbf{m}\}, \{\mathbf{r} \mapsto \mathbf{n}\}, \{\mathbf{r} \mapsto \mathbf{p}\}\}$. Our objective is to determine a single correct pair of set Θ ; i.e., select the $\{\mathbf{r} \leftrightarrow \mathbf{m}\}$

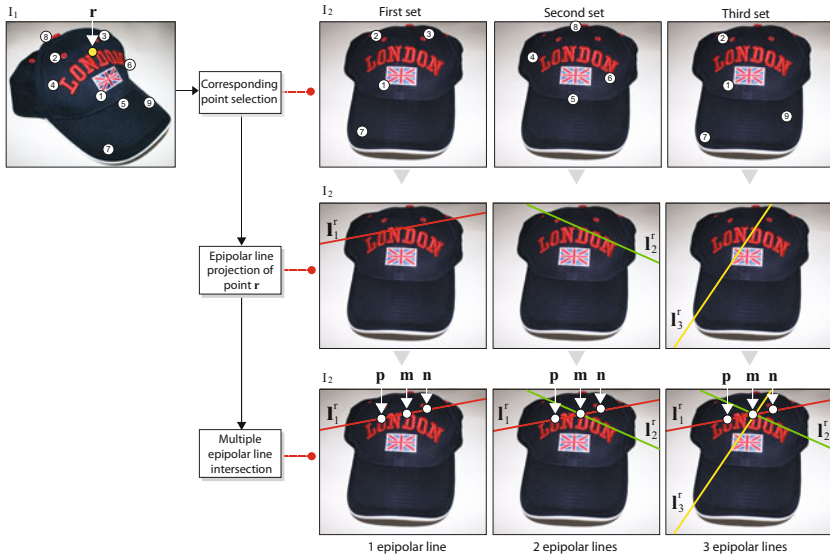


Fig. 4. New epipolar lines are created when a new set of corresponding points is created. In the example, line I_1^r was obtained with set points $\{1,2,3,7\}$, line I_2^r with set points $\{4,5,6,8\}$, and finally, line I_3^r with set points $\{1,2,7,9\}$.

pair and consequently discard the incorrect pairs. Based on the above discussion, if we intersect two epipolar lines I_1^r and I_2^r , –both generated by two different subsets of correspondences– it is clearly seen that line I_2^r is at a considerable distance from the correspondences n and p . Similarly, a third epipolar line I_3^r intersects the two previous ones at point m because the set of projected epipolar lines of point r intersect only one corresponding point in the second view, which in this case is point m , generating an *point pincl*. That effect is repeated in both images, as shown by the model of Fig. 3 and Fig. 4.

Theoretically, every new epipolar line improves the precision of corresponding point. However, in practice there is no single intersection point because of the uncalibrated nature of corresponding points used to formulate the geometric model, giving rise to an error in the estimation of the fundamental matrix. According to this analysis, one of the main problems in the estimation of the epipolar lines consists of determining their error level. Clearly, not all the epipolar lines have the same error, and for that reason we designed a method to determine the error associated with the Euclidian distance of each epipolar line. For the following analysis we will introduce the distance notation between the hypothetical point with respect to the epipolar line in the second view. Let d_k^m be the Euclidian distance between the m -th point of the second view and the epipolar line I_k^r , where r is the r -th point of the first view. The distance d_k^m is defined as

$$d_k^m = \frac{|\mathbf{m}^T \mathbf{F}_k \mathbf{r}|}{\sqrt{(\mathbf{F}_k \mathbf{r})_1^2 + (\mathbf{F}_k \mathbf{r})_2^2}}$$

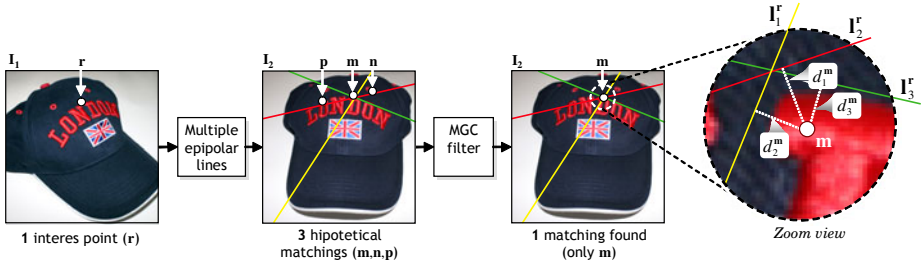


Fig. 5. The Multiple Geometric Correspondence (MGC) filter allows the determination of the point-to-point correspondence and distinguishing incorrect correspondences $\{p, n\}$

where $(\mathbf{F}_k \mathbf{r})_i$ is the i -th component of vector $\mathbf{F}_k \mathbf{r}$. As mentioned earlier, our objective is to find the correspondence $\{\mathbf{r} \leftrightarrow \mathbf{m}\}$ of the set Θ . However, we do not know the estimation error of each epipolar line. To make that estimation we will use the MLESAC algorithm, proposed by Torr and Zisserman [17]. The objective of this process is to reestimate the Euclidian distances d_1^m , d_2^m and d_3^m weighting the error of each epipolar line, a process that will be described below.

First we will introduce briefly some previous concepts of the MLESAC algorithm [17] to give greater clarity to the reader. MLESAC is a robust estimation algorithm to establish the point correspondences in multiple views, generalizing the RANSAC estimator [7]. In our proposal MLESAC is an intermediate step in the error estimation process because the error estimated by MLESAC later allows weighting the individual error of each epipolar line. One of the main advantages of MLESAC is that it is designed considering that the error $P(e)$ is a mixture of Gaussians and uniform distributions, where e is the error of the estimation of the fundamental matrix such that

$$P(e) = \left(\gamma \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{e^2}{2\sigma^2}\right) + (1 - \gamma) \frac{1}{\nu} \right) \tag{1}$$

where γ is a mixing parameter, ν is an a priori constant that indicates the distribution of the data, and σ is the standard deviation of the error in each coordinate. Parameters γ and ν are not known, but they can be estimated by means of the EM [6] algorithm. In this way, the objective function is to minimize the log-likelihood of the error, which in our case is the distance d_k^m between a point and the epipolar line, and therefore

$$-L_k = - \sum_k \left(\gamma \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp\left(-\frac{(d_k^m)^2}{2\sigma^2}\right) + (1 - \gamma) \frac{1}{\nu} \right) \tag{2}$$

We had previously mentioned that γ and ν are not known. For completeness, we now indicate how they are estimated. Assuming that there are k sets of correspondences, let η_k , where $\eta_k = 1$ if the correspondence is correct, i.e., $d_k^m = 0$, and $\eta_k = 0$ if the k correspondence is incorrect. The EM algorithm considers that η_k is an unknown value, and therefore it takes the following steps

for its estimation: (1) it generates an initial value for γ , (2) it estimates the η_k value using the initial γ estimation, and (3) it makes an estimation of γ from the new estimated value η_k , and returns to step (2). The process is repeated until it converges. For this, let p_k be the likelihood of distance d_k^m when it is an inlier, and p_o the likelihood of distance d_k^m when it is an outlier. Consequently, given the initial value of $\gamma = \frac{1}{2}$, the probabilities p_k and p_o are estimated according to

$$p_k = \gamma \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^2 \exp \left(-\frac{d_k^{m2}}{2\sigma^2} \right) \quad (3)$$

$$p_o = (1 - \gamma) \frac{1}{\nu} \quad (4)$$

Once the probabilities p_k and p_o have been estimated from the initial value γ , the following step is to reestimate the $P(\eta_k = 1|\gamma)$ value according to

$$P(\eta_k = 1|\gamma) = \frac{p_k}{p_k + p_o}, \quad (5)$$

and finally, in the phase called '*maximization*' of step (3), the value γ is reestimated according to the updated mixture of the probabilities p_k and p_o ,

$$\gamma = \frac{1}{n} \sum_k \left(\frac{p_k}{p_k + p_o} \right) \quad (6)$$

Normally three iterations are needed for the algorithm to converge. In this way the estimation of the log-likelihood of the k -th hypothesis of each epipolar line allows us to weight correctly the real distance d_k^m . To make that estimation we will use the partial values of the log-likelihood (L_k) and in that way we weight the distance d_k^m according to the following formulation:

$$\tilde{d}_k^m = d_k^m \left(\frac{|\min(L_k) - L_k| + 1}{\sum_k (L_k)} \right). \quad (7)$$

where \tilde{d}_k^m is a weighted distance that considers the error associated with each fundamental matrix. This procedure allows weighting and reestimating the distance of the epipolar lines according to the log-likelihood of the projection error with respect to the set of hypothetical points in the second view. The estimation of the error allows weighting correctly the distance d_k^m , increasing or decreasing it according to the size of its error. Therefore, to determine a correspondence, we determine the distance with respect to the set Θ . Finally, to identify the correspondence of point \mathbf{r} the following relation must be satisfied:

$$\tilde{d}_k^m < \epsilon, \quad (8)$$

where ϵ is a distance measured in pixels. The final result allows the determination of which points are corresponding and which, depending on a threshold level, must be discarded. Fig. 5 presents an example of how the error estimation discards points \mathbf{n} and \mathbf{p} from the set of correspondences Θ . In particular, the

Algorithm 1. *Bifocal Geometric Correspondence* (BIGC) algorithm in two views

- 1: Determine n sets in correspondence in two views. These sets are known or estimated in a process that can be off-line or automatic by means of the analysis of correspondences; for example, with SIFT [11] or SURF [3].
 - 2: Determine the fundamental matrix \mathbf{F}_k , for k sets in correspondence, where $k < n$.
 - 3: Determine the epipolar line \mathbf{l}_k^r of point \mathbf{r} in the first view.
 - 4: Determine the error associated with each epipolar line \mathbf{l}_k^r with the MLESAC algorithm and reestimate the real distance \tilde{d}_k^m between the hypothetical correspondence and the epipolar line.
 - 5: Assign the correspondence to point \mathbf{m} provided that the restriction $\tilde{d}_k^m < \epsilon$ is fulfilled for all $\mathbf{m} \in \Theta$.
-

Multiple Geometric Correspondence MGC filter' block is in charge of reestimating the distances (Fig. 5). In the example, once point \mathbf{m} is chosen, only the $\{\mathbf{r} \leftrightarrow \mathbf{m}\}$ combination is possible.

As shown in the previous steps, in spite of the errors existing in the estimation of the epipolar lines, their set allows the estimation of point-to-point correspondence. A complete description of the proposed methodology, which we call *Bifocal Geometric Correspondence* (BIGC), is presented in the 1 Algorithm.

3 Experimental Results

This section presents the experimental results generated with sequences of uncalibrated images in two views. A set of 10 stereo images composed mainly of landscapes and walls, most of them supplied by the authors (Fig. 6) were used. In all the experiments we have considered two standard indicators [13]: $r = \frac{TP}{TP+FN}$ (recall) and $p = \frac{TP}{TP+FP}$ (precision). TP is the number of *true positives* or correctly classified correspondences. FN is the number of *false negatives* or real correspondences not detected by our algorithm. FP is the number of *false positives* or correspondences classified incorrectly. These two indicators can be joined in a single measure F-score = $\frac{2 \cdot p \cdot r}{p+r}$ [13]. Ideally, one can expect that $r = 100\%$, $p = 100\%$, and F-score = 1.

The first test set is composed of 10 pairs of images with a resolution of 1200 \times 800 pixels. The main existing geometric transformations are perspective, rotation, translation, and different degree scale (Fig. 6). This set consists of landscapes and walls in settings with natural lighting, showing a large number of regions in correspondence. According to the steps described in the 1 Algorithm, the first step consists of determining n sets of corresponding pairs. This process was performed with the SURF [3] algorithm, from which we selected the best k sets with the least projection error according to the MLESAC estimator. To evaluate the performance of the algorithm we determined 300 corresponding points in random positions within each pair of images in a process carried out off-line by means of the SURF algorithm. We then evaluated the ability of the algorithm to determine the correspondence by varying the $k \in [1, \dots, 14]$ parameter and

the $\epsilon \in [0, \dots, 10]$ parameter. Note that in the latter parameter the values are rounded.

Below we present the results according to the variations of the k and ϵ parameters. In the first case we analyze the influence of the k parameter keeping distance ϵ fixed. As seen in Fig. 7a, our best performance had an F-score= 0.97 at a discretized distance $\epsilon = 0$, using the intersection of three fundamental matrices ($k = 3$). It is interesting to mention that as the ϵ parameter increases, performance starts dropping. This indicates that the method is very precise in these kinds of images because there is a large number of correspondences. In the second case we analyze the influence of parameter ϵ keeping fixed the number of solutions k . According to the results obtained, we see a maximum performance at $k = 3$. On the contrary, an increase of this value decreases the performance of the algorithm because the projection error increases.

Remember that in the analyzed sequence there is a large number of correspondences in spite of the geometric transformations present in them. Therefore, these results indicate that it is possible to use and estimate geometric models in two images with high precision at a subpixel resolution. According to the performance indicated in Fig. 7a, after $k = 4$ there is no improvement in the performance for $\epsilon > 4$.

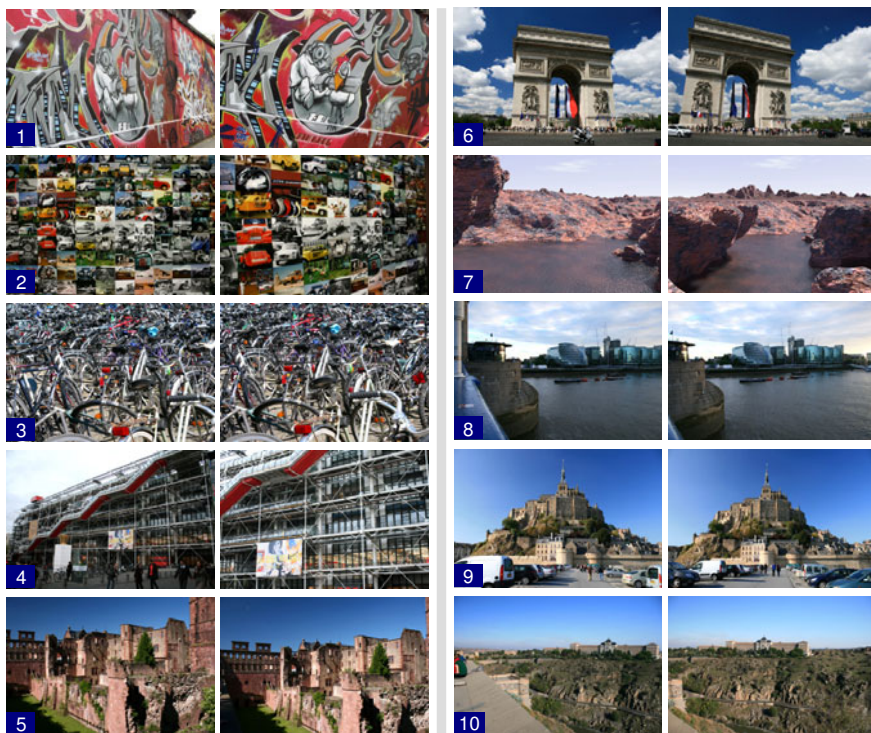


Fig. 6. Outdoor set of 10 stereo images

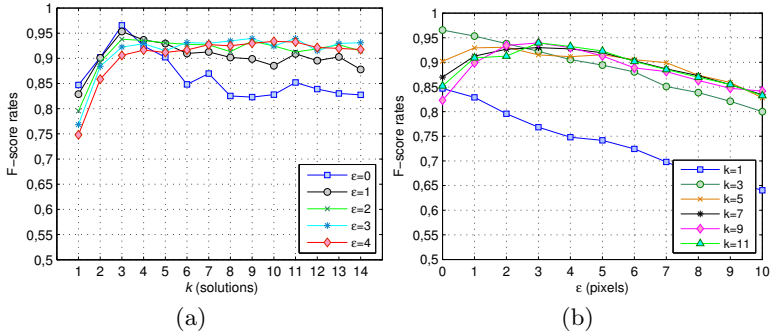


Fig. 7. Average performance of the set of outdoor images (a) Influence of the k parameter as the maximum tolerance distance in pixels ϵ varies (b) Influence of distance (ϵ) on the detection of correspondences for different numbers of solutions (k)

4 Conclusions

In this paper we have developed two important contributions. First, we presented a method that uses the intersection of multiple geometric solutions in two views and in three views to determine point-to-point correspondence. Second, for each geometric model we have determined the real distance with respect to corresponding point by means of the MLESAC estimator. The main novelty of our proposal is the geometric methodology for solving the problem of the estimation of point-to-point correspondence, regardless of the angles of the points of view of the objects. We call this algorithm Bifocal Geometric Correspondence (BIGC) for the correspondence in two views.

It is important to note that the point can be occluded, but its position remains valid because our method is based on a geometric model that defines the scene. We also show that the use of multiple random solutions makes it possible to improve the performance of the correspondence in two views. Although our method starts from the basis that there is a set of points in previous correspondence necessary to determine the fundamental matrices, it is designed to maximize the correspondences in specific regions of each image.

In the experiments performed we considered outdoor images. The results obtained with these sets indicate that the BIGC algorithm was capable of determining point-to-point correspondence precisely, with a performance F-score= 97% in stereo images at a discretized distance $\epsilon = 0$ pixels for outdoor images. For all the images analyzed, we showed that the point-to-point correspondence can be generated through a multiple geometric relation between two views. In relation to this last point we stress that our method can be applied as support of industrial control to follow-up faults in uncalibrated sequences, among other applications.

Acknowledgment. This work was supported by the National Commission of Science and Technology (CONICYT, Chile). Fondecyt grant no. 11100098.

References

1. Barron, J.L., Fleet, D.J., Beauchemin, S.S.: Performance of optical flow techniques. *International Journal of Computer Vision* 12(1), 43–77 (1994)
2. Bartoli, A., Sturm, P.: Nonlinear estimation of the fundamental matrix with minimal parameters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(3), 426–432 (2004)
3. Bay, H., Ess, A., Tuytelaars, T., Gool, L.: Surf: Speeded up robust features. *Computer Vision and Image Understanding (CVIU)* 110(3), 346–359 (2008)
4. Caspi, Y., Irani, M.: A step towards sequence-to-sequence alignment. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 682–689. IEEE, Hilton Head Island (2000)
5. Caspi, Y., Simakov, D., Irani, M.: Feature-based sequence-to-sequence matching. *International Journal of Computer Vision* 68(1) (2006)
6. Dempster, A.P., Laird, N., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B* 39, 1–38 (1977)
7. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM* 24(6), 381–395 (1981)
8. Fitzgibbon, A.: Robust registration of 2d and 3d point sets. *Image and Vision Computing* 21(13-14) (December 2003)
9. Hartley, R., Zisserman, A.: *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge (2000)
10. Kadir, T., Zisserman, A., Brady, M.: An Affine Invariant Salient Region Detector. In: Pajdla, T., Matas, J. (eds.) *ECCV 2004*. LNCS, vol. 3021, pp. 228–241. Springer, Heidelberg (2004)
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60(2), 91–110 (2004)
12. Moreels, P., Perona, P.: Evaluation of features detectors and descriptors based on 3d objects. *International Journal of Computer Vision* 73(3) (2007)
13. Olson, D.L., Delen, D.: *Advanced Data Mining Techniques*. Springer, Heidelberg (2008)
14. Romano, R.: *Projective Minimal Analysis of Camera Geometry*. Phd. thesis, M.I.T., USA (May 2002)
15. Tordoff, B.J., Murray, D.W.: Guided-mlesac: faster image transform estimation by using matching priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(10), 1523–1535 (2005)
16. Torr, P.: Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *International Journal of Computer Vision* 50(1), 35–61 (2002)
17. Torr, P., Zisserman, A.: Mlesac: A new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding* 78, 138–156 (2000)
18. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: A survey. *Computer Graphics and Vision* 3(3), 177–280 (2007)
19. Vidal, R., Ma, Y., Soatto, S., Sastry, S.: Two-view multibody structure from motion. *International Journal of Computer Vision* 68(1) (June 2006)
20. Zhang, Z., Deriche, R., Faugeras, O., Luong, Q.T.: A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence* 78(1-2), 87–119 (1995)

Digital Hologram Compression Using Correlation of Reconstructed Object Images

Jae-Young Sim

School of Electrical and Computer Engineering,
Ulsan National Institute of Science and Technology,
Ulsan 689-798, South Korea
jysim@unist.ac.kr

Abstract. An efficient digital hologram compression algorithm is proposed using the correlation in the complex valued object image. While the pure values are almost uncorrelated, the magnitude values exhibit a strong correlation between the real and imaginary part object images. Therefore, we adaptively employ the encoding result of one image to encode another image. Both images are first wavelet transformed and the wavelet coefficients are encoded using the SPIHT method. We used the significance encoding result of the real part image as the contexts of arithmetic coder for encoding the imaginary part image. Experimental results demonstrate that the proposed algorithm yields a better compression performance than the conventional method.

Keywords: Digital hologram, digital hologram compression, context-adaptive arithmetic coding.

1 Introduction

Three-dimensional (3D) images facilitate more realistic and immersive visual experiences, and therefore have drawn much attention in recent years. Holography is considered as one of the most promising techniques for 3D image representation, since it is free from eye fatigue and viewpoint constraints. Fig. 1 shows a typical principle of holography [7]. A light wave is split into two parts such that one directly travels to the recording medium as a reference light, and the other illuminates a 3D object. The interference pattern between the reference light and the reflected light on the 3D object is recorded, which is called hologram. When the hologram is illuminated with the reference light, the virtual object image is reconstructed.

With the aid of high resolution image sensors and high performance computers, holograms can be stored and processed in digital form [8,13,6,12]. To be specific, an interference pattern is captured by a CCD (charged coupled device) sensor and stored as a digital image which is called digital hologram. In addition, the object image can be numerically calculated with computers based on the mathematical modeling of the reconstruction process [12]. The numerically reconstructed image yields the artifacts such as the bright region of undiffracted

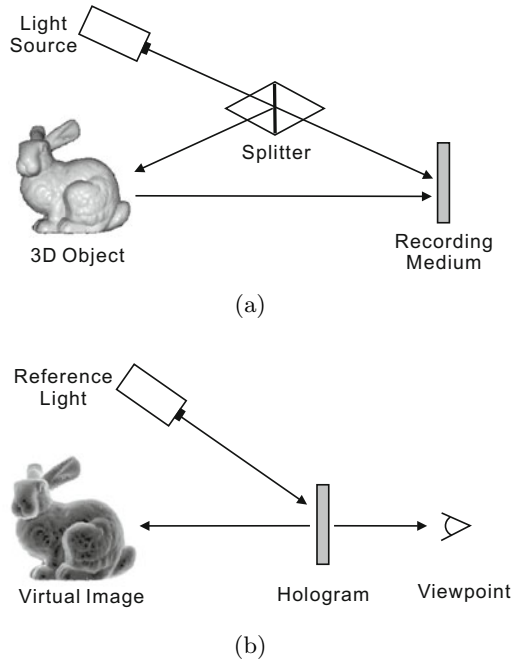


Fig. 1. Principle of holography. (a) Hologram recording and (b) image reconstruction

light wave and the conjugate object image. To remove such drawbacks, phase-shifting digital holography was developed which employs multiple interference patterns to directly extract the complex amplitude of the object wave [17].

In order to sufficiently sample the detailed shape of interference pattern, digital hologram is generally stored as a high resolution image which requires a huge amount of storage space. Therefore, a lot of research has been performed to efficiently compress the digital holograms [10,15,16,4,5,3]. Various entropy coding methods were used to losslessly compress the digital holograms [10]. Shortt *et al.* transformed the phase-shifting digital holograms using wavelets and encoded the quantized wavelet coefficients [15]. A nonuniform quantization method was also proposed based on the histogram of digital hologram [16]. While the most algorithms have directly compressed the digital hologram data, Darakis and Soraghan developed the compression algorithms for reconstructed images [4,5,3]. The wavelet-like Fresnelet transform is used and the transformed coefficients of the reconstructed image are encoded [4]. The compression performance of the digital hologram was also compared to that of the reconstructed image [5,3]. Note that, progressive transmission can be facilitated in [5] and [3], since the embedded bitstreams are derived by using the SPIHT (set partitioning in hierarchical trees) method [11].

However, the previous compression techniques did not exploit the correlation in the complex valued reconstructed wave field from a digital hologram. In this paper, we investigate the correlation between the real and imaginary part data of

the complex valued object image, and propose an efficient compression algorithm for digital holograms. It is observed that the magnitude images of the real and imaginary parts are highly dependent. Therefore we use the encoding result of the real part image as contexts to adaptively encode the imaginary part image based on the context-adaptive arithmetic coding. Experimental results demonstrate that the proposed algorithm exhibits a better rate-distortion performance than the existing method to compress digital holograms.

This paper is organized as follows. Section 2 describes the digital holography. The correlation of digital hologram is investigated in Section 3. Section 4 explains the coding algorithm of digital hologram data. Section 5 presents the experimental results, and finally Section 6 concludes this paper.

2 Digital Holography

Fig. 2 illustrates a coordinate system in digital holography, where the digital hologram is captured in the hologram plane (or camera plane) and the reconstructed image is calculated in the plane of distance d from the hologram plane. Let $\mathcal{U}(x, y)$ be the complex amplitude of the reflected object wave in the hologram plane, which is represented by

$$\mathcal{U}(x, y) = a_u(x, y)\exp\{i\phi_u(x, y)\},$$

where $a_u(x, y)$ and $\phi_u(x, y)$ denote the real amplitude and the phase, respectively. Similarly, the reference light wave is given by

$$\mathcal{V}(x, y) = a_v(x, y)\exp\{i\phi_v(x, y)\}.$$

The intensity of the interference pattern of $\mathcal{U}(x, y)$ and $\mathcal{V}(x, y)$ is the scaled version of digital hologram, which is calculated as

$$I(x, y) = |\mathcal{U}(x, y) + \mathcal{V}(x, y)|^2$$

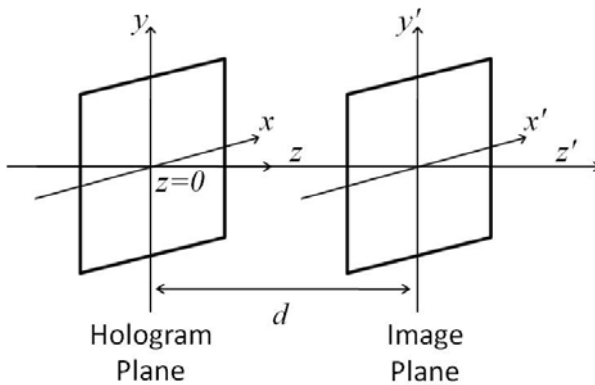


Fig. 2. Coordinate system in digital holography

$$\begin{aligned}
&= (\mathcal{U}(x, y) + \mathcal{V}(x, y)) (\overline{\mathcal{U}(x, y) + \mathcal{V}(x, y)})^* \\
&= a_u^2(x, y) + a_v^2(x, y) + \mathcal{U}(x, y)\mathcal{V}^*(x, y) + \mathcal{V}(x, y)\mathcal{U}^*(x, y), \quad (1)
\end{aligned}$$

where $*$ means the complex conjugate.

When reconstructing an image from the digital hologram, $I(x, y)$ is multiplied by the reference wave as

$$\mathcal{V}(x, y)I(x, y) = (a_v^2(x, y) + a_u^2(x, y)) \mathcal{V}(x, y) + a_v^2(x, y)\mathcal{U}(x, y) + \mathcal{V}^2(x, y)\mathcal{U}^*(x, y).$$

The first term denotes the undiffracted component passing through the hologram. $a_v^2(x, y)\mathcal{U}(x, y)$ and $\mathcal{V}^2(x, y)\mathcal{U}^*(x, y)$ mean the virtual object image and the distorted real object image, respectively. In practice, as shown in Fig. 2, the complex amplitude $\mathcal{W}(x', y')$ in the image plane is reconstructed by propagating $\mathcal{V}(x, y)I(x, y)$ from the hologram plane to the image plane via the Fresnel-Kirchhoff integral in (2), which models the diffraction of light wave [14,12].

$$\mathcal{W}(x', y') = \frac{i}{\lambda} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \mathcal{V}(x, y)I(x, y) \frac{\exp(-i\frac{2\pi}{\lambda}D)}{D} dx dy, \quad (2)$$

where λ is the wavelength of the reference light and $D = \sqrt{(x - x')^2 + (y - y')^2 + d^2}$. Equation (2) is numerically approximated by

$$\begin{aligned}
\mathcal{W}(x', y') &= \frac{i}{\lambda d} \exp\left(-i\frac{2\pi d}{\lambda}\right) \exp\left\{-i\frac{\pi}{\lambda d}(x'^2 + y'^2)\right\} \\
&\times \iint_{-\infty}^{\infty} \mathcal{V}(x, y)I(x, y) \exp\left\{\frac{-i\pi}{\lambda d}(x^2 + y^2)\right\} \exp\left\{\frac{i2\pi}{\lambda d}(xx' + yy')\right\} dx dy, \quad (3)
\end{aligned}$$

which is called Fresnel transform. Moreover, the digitized version of Fresnel transform can be efficiently calculated by using the fast Fourier transform [14,12].

3 Correlation in Digital Hologram

The complex valued image $\mathcal{W}(x', y')$ of 3D object is reconstructed from a digital hologram using the diffraction of light waves via (2) or (3). Therefore, $\mathcal{W}(x', y')$ is composed of the real part image $Re(\mathcal{W}(x', y'))$ and the imaginary part image $Im(\mathcal{W}(x', y'))$,

$$\mathcal{W}(x', y') = Re(\mathcal{W}(x', y')) + iIm(\mathcal{W}(x', y')).$$

Fig. 3 (a) represents the ‘Brahms’ hologram. Fig. 3 (b) and (c) show the magnitude images of the corresponding $Re(\mathcal{W}(x', y'))$ and $Im(\mathcal{W}(x', y'))$, respectively. While the digital hologram yields a noise-like pattern and thus does not directly represent the shape of captured object, the reconstructed image exhibits a relatively higher spatial correlation by showing the shape of object. Therefore, a more improved compression performance can be achieved on the reconstructed images than the digital holograms [5,3]. Moreover, as observed in Fig. 3 (b) and

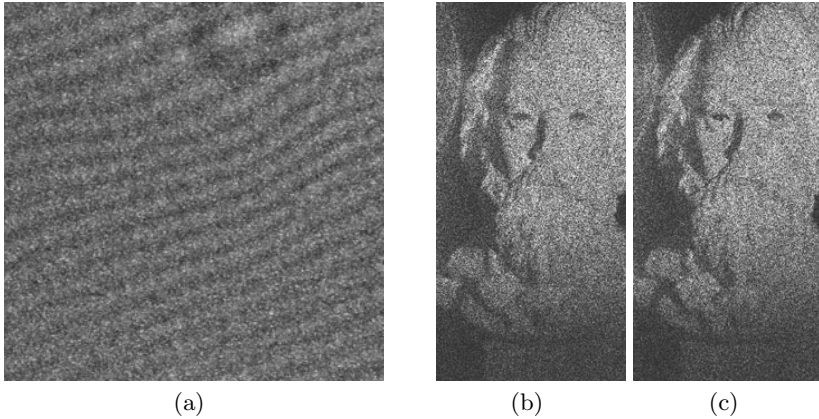


Fig. 3. Digital hologram and the reconstructed object images. (a) The ‘Brahms’ digital hologram. (b) Magnitude of the real part object image. (c) Magnitude of the imaginary part object image.

(c), the real and imaginary part images also yield the similar magnitude values each other, at the corresponding pixels.

In order to quantitatively investigate the dependence between $Re(\mathcal{W}(x', y'))$ and $Im(\mathcal{W}(x', y'))$, we measure the Pearson correlation coefficient

$$\rho(R, I) = \frac{E[R - E[R]] E[I - E[I]]}{\sqrt{E[(R - E[R])^2] E[(I - E[I])^2]}} \quad (4)$$

where R and I are the random variables of $Re(\mathcal{W}(x', y'))$ and $Im(\mathcal{W}(x', y'))$, respectively. In general, $\rho(R, I) \approx 0$. It means that the pure values of $Re(\mathcal{W}(x', y'))$ and $Im(\mathcal{W}(x', y'))$ are almost uncorrelated each other, yielding different patterns of speckles. However, it is empirically observed that the magnitude values of $|Re(\mathcal{W}(x', y'))|$ and $|Im(\mathcal{W}(x', y'))|$ have a relatively strong correlation, for example 0.6 in the ‘Brahms’ hologram.

The similar characteristics are also observed in the wavelet transformed images. Fig. 4 (a) and (b) represent the magnitude images of the wavelet coefficients of $Re(\mathcal{W}(x', y'))$ and $Im(\mathcal{W}(x', y'))$, respectively, which are reconstructed from the ‘Copper screw’ hologram. The correlation coefficient for the wavelet coefficients is 0.53.

4 Digital Hologram Compression

4.1 Progressive Coding

One of the state-of-the-art compression algorithms of digital holograms was developed on the reconstructed images using the wavelet-like transform and the SPIHT coding method [3]. We also apply the wavelet transform to the real part and the imaginary part of the complex valued reconstructed image individually,

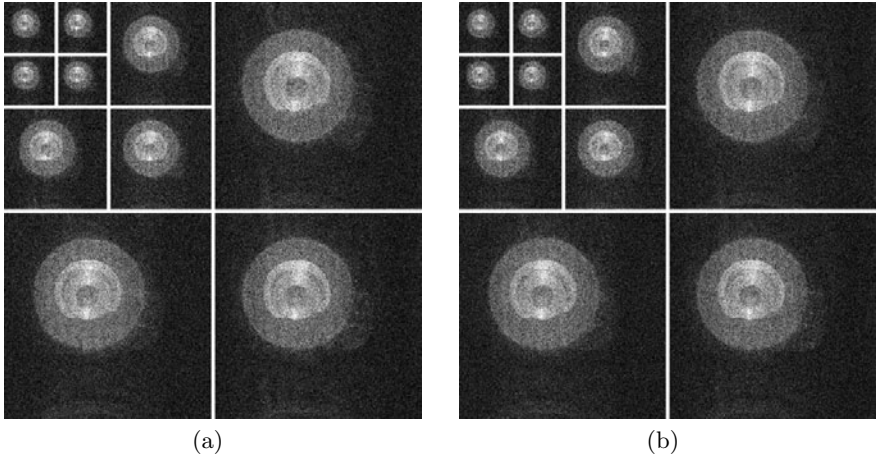


Fig. 4. Magnitude of wavelet coefficients in the object image reconstructed from the ‘Copper screw’ digital hologram. (a) Real part image. (b) Imaginary part image.

using the 9/7 filters [2]. The wavelet coefficients are uniformly quantized with the same step size, and then progressively encoded using the SPIHT method [11]. As shown in Fig. 4, the reconstructed image of a digital hologram is relatively sparse in the sense of significant pixels when compared with the general images, since it represents the sectional image of 3D objects focused on a specific depth. As a result, a coding gain can be expected by encoding a group of pixels together in the reconstructed image.

4.2 Context-Adaptive Coding

For each image of the real and imaginary parts, bitplanes of the quantized wavelet coefficients are progressively encoded one by one from the most significant bitplane (MSB) to the least one based on the SPIHT method. Let $c(x, y)$ be the quantized wavelet coefficient at the pixel coordinate (x, y) . For a given b th MSB, the significance coding is performed. If $c(x, y)$ first becomes significant, i.e. $2^b \leq |c(x, y)| < 2^{b+1}$, then the significance bit $S_b(c(x, y)) = 1$ is encoded and the sign bit is encoded. If $c(x, y)$ is insignificant as $|c(x, y)| < 2^b$, then $S_b(c(x, y)) = 0$ is encoded. When $c(x, y)$ is already significant at the previous bitplane such that $|c(x, y)| \geq 2^{b+1}$, the magnitude refinement bit is encoded. The resulting bit sequence is entropy encoded using the arithmetic coder [9].

The significance of wavelet coefficient is directly related to the magnitude which yields a strong correlation between the real and imaginary part images. It means that if the coefficient $c_{real}(x, y)$ in the real part image is significant (or insignificant) at a given bitplane, the coefficient $c_{imag}(x, y)$ in the imaginary part image is highly probable to be significant (or insignificant) at the same bitplane. Therefore, the result of significance coding for one image can be used to predict the significance of another image. In practice, $S_b(c_{real}(x, y))$ is first initialized to be ‘0’ for all b , and then updated while encoding $c_{real}(x, y)$. When encoding the

imaginary part image, if $S_b(c_{real}(x, y)) = 1$, then $S_b(c_{imag}(x, y))$ is arithmetic encoded with the context A_{pixel} , otherwise with the context B_{pixel} . Furthermore, we employ the different contexts according to the frequency subbands of wavelet decomposition, since the correlation coefficient between the real and imaginary part images exhibits different characteristics in each subband.

In addition, the SPIHT method improves the coding gain by encoding the significance of a group of coefficients together. Let $G(x, y)$ be the group of all pixel coefficients belonging to the hierarchical tree derived from the coordinate (x, y) . If all the coefficients in $G(x, y)$ are insignificant at the b th MSB, then the significance bit $S_b(G(x, y)) = 0$ is encoded. Otherwise $S_b(G(x, y)) = 1$ is encoded, and the group $G(x, y)$ is divided into the smaller sub-groups. The significance coding is iteratively performed on each sub-group until the coordinates of significant coefficients are encoded. The group $G_{real}(x, y)$ in the real part image and the group $G_{imag}(x, y)$ in the imaginary part image are derived from the same spatial region, and thus yield the similar characteristics of significance. Therefore, we employ the other contexts A_{group} and B_{group} to adaptively encode $S_b(G_{imag}(x, y))$, according to the result of $S_b(G_{real}(x, y))$ which is updated during the encoding of the real part image.

5 Experimental Results

The performance of the proposed algorithm is evaluated using the three digital holograms, ‘Brahms’, ‘Dice’, and ‘Copper screw’, whose reconstructed images are shown in Fig. 5. Table 1 provides the capturing conditions for these digital holograms. The distortion of the reconstructed image is measured by the normalized root mean square (NRMS) error [3], and given by

$$\sqrt{\frac{\sum_x \sum_y (|\mathcal{W}(x, y)|^2 - |\tilde{\mathcal{W}}(x, y)|^2)^2}{\sum_x \sum_y (|\mathcal{W}(x, y)|^2)^2}} \quad (5)$$

where $\mathcal{W}(x, y)$ and $\tilde{\mathcal{W}}(x, y)$ denote the original image and the decoded one from the compressed bitstream, respectively.

Fig. 6, Fig. 7, and Fig. 8 show the rate and distortion compression performances for the imaginary part object images of the ‘Brahms’, ‘Dice’, and ‘Copper screw’ holograms, respectively. The dashed lines exhibit the results of the conventional coding method [3] and the solid lines represent the results of the proposed context-adaptive coding method. In order to compare the performances from only the coding schemes, we also used the 9/7 filters [2] to get the wavelet coefficients in the conventional method. It is observed that the proposed compression algorithm provides a better rate-distortion performance compared with the conventional method which separately encodes the real and imaginary part images. The bitrate reduction is over 20% especially at low bitrates. For example, while the conventional method consumes 11417 bytes to encode the imaginary part image to provide the distortion 7.65×10^{-2} for the ‘Brahms’ hologram, the proposed algorithm only requires 8863 bytes to achieve the same distortion.

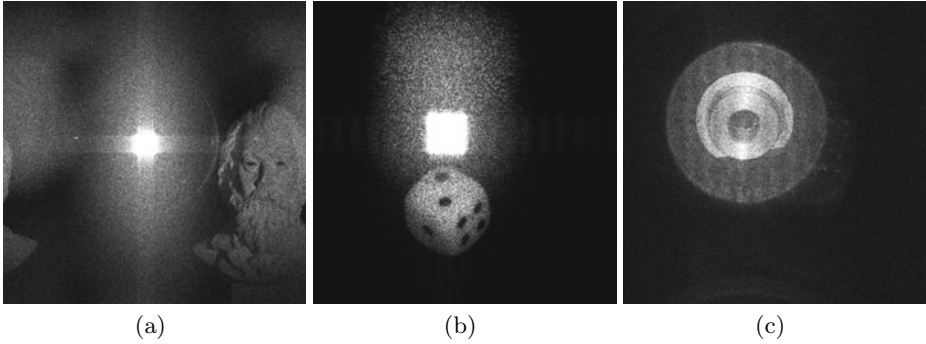


Fig. 5. Reconstructed images of the test digital holograms. (a) Brahms (Courtesy of the HoloVision project [1]). (b) Dice (Courtesy of Dr. U. Schnars). (c) Copper screw (Courtesy of Dr. F. Zhang). The ‘Brahms’ and the ‘Dice’ are the classic digital holograms and the ‘Copper screw’ is the phase-shifting digital hologram.

Table 1. Capturing conditions of digital holograms

	Brahms	Dice	Copper screw
Type	classic	classic	phase shifting
Image resolution	1024×1024	1024×1024	1024×1024
Pixel size	$6.8 \mu\text{m}$	$6.8 \mu\text{m}$	$6.45 \mu\text{m}$
Wavelength of reference light	632.8 nm	632.8 nm	830 nm
Captured distance	1290 mm	1054 mm	285 mm

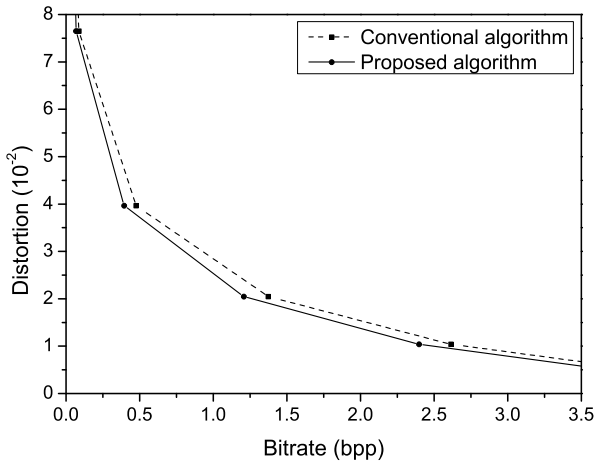


Fig. 6. Rate and distortion curve of the proposed algorithm compared with the conventional method. The rate means the average bits per pixel to encode the imaginary part object image reconstructed from the ‘Brahms’ digital hologram.

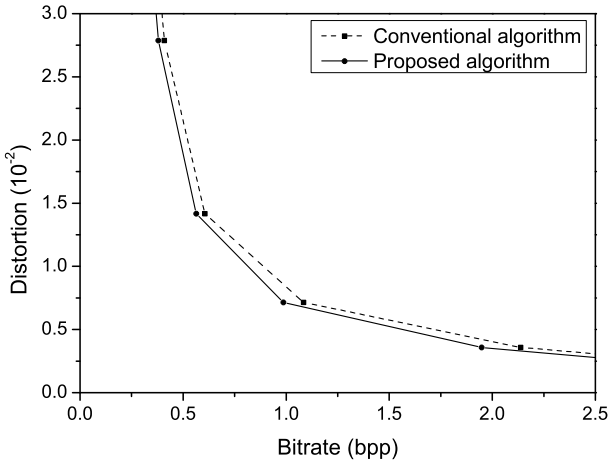


Fig. 7. Rate and distortion curve of the proposed algorithm compared with the conventional method. The rate means the average bits per pixel to encode the imaginary part object image reconstructed from the ‘Dice’ digital hologram.

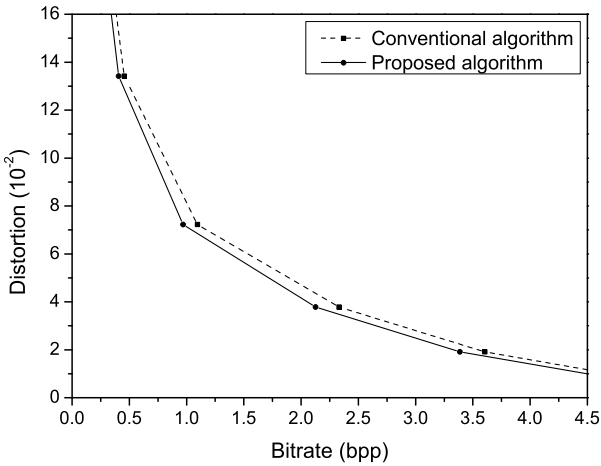


Fig. 8. Rate and distortion curve of the proposed algorithm compared with the conventional method. The rate means the average bits per pixel to encode the imaginary part object image reconstructed from the ‘Copper screw’ digital hologram.

6 Conclusion

In this paper, we proposed an adaptive compression algorithm for digital holograms using the correlation in the reconstructed images. The object image reconstructed from a digital hologram is inherently a complex wave field, and thus composed of the real and imaginary parts. The real and imaginary part images are wavelet transformed and the transform coefficients are progressively encoded using the SPIHT method. The two images yield a relatively strong correlation each other, which can be exploited to improve the compression performance for digital holograms. Specifically, the result of significance coding for the real part image is employed as contexts of arithmetic coder to encode the imaginary part image. Experimental results showed that the proposed context-adaptive compression algorithm for digital holograms improved the coding gain compared with the conventional state-of-the-art method.

Acknowledgments. This research was supported by Basic Science Research Program through the NRF of Korea funded by the Ministry of Education, Science and Technology (2010-0006595).

References

1. <http://www.edge.no/projects/holovision/>
2. Antonini, M., Barlaud, M., Mathieu, P., Daubechies, I.: Image coding using wavelet transform. *IEEE Trans. Image Processing* 1(2), 205–220 (1992)
3. Darakis, E., Naughton, T.J., Soraghan, J.J.: Compression defects in different reconstructions from phase-shifting digital holographic data. *Appl. Opt.* 46(21), 4579–4586 (2007)
4. Darakis, E., Soraghan, J.J.: Use of Fresnelets for phase-shifting digital hologram compression. *IEEE Trans. Image Processing* 15(12), 3804–3811 (2006)
5. Darakis, E., Soraghan, J.J.: Reconstruction domain compression of phase-shifting digital holograms. *Appl. Opt.* 46(3), 351–356 (2007)
6. Frauel, Y., Naughton, T.J., Matoba, O., Tajahuerce, E., Javidi, B.: Three-dimensional imaging and processing using computational holographic imaging. *Proc. of IEEE* 94(3), 636–653 (2006)
7. Gabor, D.: A new microscopic principle. *Nature* 161, 777–778 (1948)
8. Goodman, J.W., Lawrence, R.W.: Digital image formation from electronically detected holograms. *Appl. Phys. Lett.* 11, 77–79 (1967)
9. Moffat, A., Neal, R., Witten, I.H.: Arithmetic coding revisited. In: *Proc. IEEE Data Compression Conference*, pp. 202–211 (March 1995)
10. Naughton, T.J., Frauel, Y., Javidi, B., Tajahuerce, E.: Compression of digital holograms for three-dimensional object reconstruction and recognition. *Appl. Opt.* 41, 4124–4132 (2002)
11. Said, A., Pearlman, W.: A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans. Circuits Syst. Video Technol.* 6, 243–250 (1996)
12. Schnars, U., Jueptner, W.: *Digital holography: digital hologram recording, numerical reconstruction, and related techniques*. Springer, Heidelberg (2005)

13. Schnars, U., Jüptner, W.: Direct recording of holograms by a CCD target and numerical reconstruction. *Appl. Opt.* 33, 179–181 (1994)
14. Schnars, U., Jüptner, W.: Digital recording and numerical reconstruction of holograms. *Meas. Sci. Technol.* 13, R81–R101 (2002)
15. Shortt, A.E., Naughton, T.J., Javidi, B.: Compression of digital holograms of three-dimensional objects using wavelets. *Opt. Expr.* 14(7), 2625–2630 (2006)
16. Shortt, A.E., Naughton, T.J., Javidi, B.: Histogram approaches for lossy compression of digital holograms of three-dimensional objects. *IEEE Trans. Image Processing* 16(6), 1548–1556 (2007)
17. Yamaguchi, I., Zhang, T.: Phase-shifting digital holography. *Opt. Lett.* 13(9), 1268–1270 (1997)

Pedestrian Image Segmentation via Shape-Prior Constrained Random Walks

Ke-Chun Li, Hong-Ren Su, and Shang-Hong Lai

Department of Computer Science,
National Tsing Hua University Hsinchu, Taiwan

Abstract. In this paper, we present an automatic and accurate pedestrian segmentation algorithm by incorporating pedestrian shape prior into the random walks segmentation algorithm. The random walks [1] algorithm requires user-specified labels to produce segmentation with each pixel assigned to a label, and it can provide satisfactory segmentation result with proper input labeled seeds. To take advantage of this interactive segmentation algorithm, we improve the random walks segmentation algorithm by incorporating prior shape information into the same optimization formulation. By using the human shape prior, we develop a fully automatic pedestrian image segmentation algorithm. Our experimental results demonstrate that the proposed algorithm significantly outperforms the previous segmentation methods in terms of pedestrian segmentation accuracy on a number of real images.

Keywords: human segmentation, random walks, shape prior.

1 Introduction

Pedestrian segmentation is an important problem in computer vision, especially for video surveillance [2]. Human detection is usually the first step in video surveillance. The traditional systems can only supply rough human locations, but more precise human segmentation information is required for some advanced applications, such as gait recognition, human identification or human motion analysis [3]. Therefore, pedestrian image segmentation which segments pedestrian from an image is a critical problem with several potential applications.

Object segmentation is the key technique in many applications, including interactive video editing, content-based image retrieval, video surveillance, medical image analysis, and so on. This problem has been researched in computer vision for decades, but automatic object segmentation is still very challenging for general objects whose appearance is difficult to model, such as humans. A finely segmented human image can provide important and precise information of the human, which is very helpful for a number of higher level tasks on human motion analysis. It remains a great challenge because of the highly articulated human body postures, viewpoint changes, large appearance variations, and cluttered background, especially when pedestrians have similar color or texture with

the connected background regions. Recently, interactive segmentation techniques [1][20] becomes popular due to their flexibility in handling the difficult cases.

In recent years, simultaneous detection and segmentation of pedestrians become a popular problem. In general, these methods obtain rough shape or silhouette of human from different ways. In [4], Lin et al. proposed a hierarchical part-template matching approach [5] and learned a human detector which consists of elementary part detectors for head-torso, upper legs, and lower legs. This algorithm provides accurate human detection and a rough human segmentation from an image. Gao et al. [6] presented a novel feature representation called Adaptive Contour Feature (ACF) that is robust against reasonable object deformation like HOG, and the detection and segmentation of human was trained by a cascade framework [7] and Real AdaBoost [8]. For human segmentation, each pixel is classified as human or background from the ACFs.

The random walker algorithm was originally developed for interactive image segmentation by Grady [1]. With input of some user-specified seeds, the algorithm provides image segmentation result by solving a sparse linear system. Grady [21] extended the random walker algorithm by incorporating an intensity-based prior model into the energy minimization formulation to relieve the requirement of user-specified labels in the original algorithm. The prior model proposed in [21] is based on imposing the prior intensity distribution constraint as an additional quadratic energy term into the original energy minimization formulation. In this work, we propose to impose a pedestrian shape prior, instead of the intensity-based prior, into the random walks image segmentation algorithm. This shape prior is a mixture of Gaussians distribution, which is quite different from [21], and the image segmentation can still be obtained via solving similar sparse linear systems.

The overall flow of the proposed pedestrian segmentation algorithm is depicted in Fig. 1. It mainly consists of two components; namely, shape prior estimation and pedestrian segmentation. The shape prior estimation is to learn the human shape prior distribution from a training set of human shapes. The pedestrian segmentation is based on a shape-prior constrained random walks

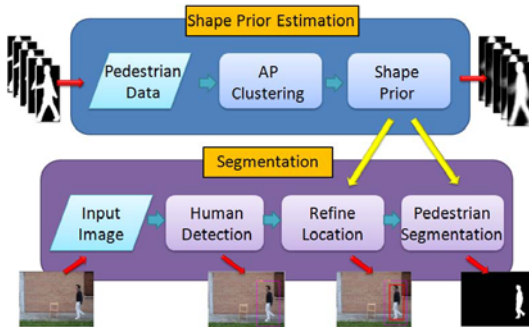


Fig. 1. System flow of the proposed pedestrian segmentation algorithm

image segmentation algorithm that incorporates the learned human shape prior into the random walks segmentation framework. We will detail these two main components in the subsequent sections.

2 Shape Prior Estimation

In order to estimate the pedestrian shape prior model, we collect a large set of pedestrian silhouettes, which were extracted from videos with uniform background. We took several pedestrian sequences, with each sequence containing one person walking along different directions. We applied a background subtraction procedure to segment the pedestrian regions from videos and the extracted regions are normalized to 64×128 , as depicted in Fig. 2. Finally, the dataset consists of 1,439 pedestrian silhouettes extracted from videos of 10 walking persons. With the left-right reflections, there are totally 2,878 pedestrian silhouettes in the training dataset.

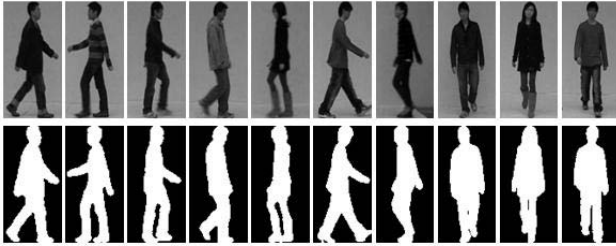


Fig. 2. Samples of pedestrian silhouettes extracted from images. The silhouettes are normalized to 64×128 .

2.1 Pedestrian Shape Prior Model

The prior pedestrian shape model is estimated from the binary pedestrian silhouette data. To cluster the set of human silhouette data, we employ the Affinity Propagation (AP) [9] clustering algorithm in this paper. The AP clustering is an iterative algorithm that works by finding a set of exemplars in the data and assigning other data points to the exemplars.

In this work, all the training pedestrian silhouettes are divided into 7 clusters after applying the AP clustering. Next, we estimate the shape prior model $\boldsymbol{\mu}^s = (\mu_1^s, \mu_2^s, \dots, \mu_N^s)$ by taking the averages for all pixels of the silhouettes for each cluster. In addition to the 7 prior models constructed from all the associated pedestrian silhouettes, we also compute the model with all the pedestrian images $\boldsymbol{\mu}^0 = (\mu_1^0, \mu_2^0, \dots, \mu_N^0)$. Fig. 3. shows the probability map of each cluster. All the 8 shape models are employed to form a mixture of Gaussians distribution for the human shape prior distribution; namely,

$$p(\mathbf{x}) = c \sum_{s=0}^7 G(\mathbf{x} - \boldsymbol{\mu}^s, D^s), \quad (1)$$

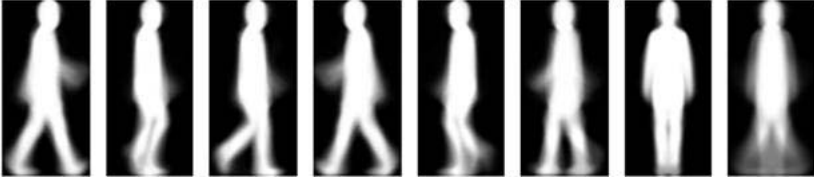


Fig. 3. The probability maps corresponding to all clusters are shown here, and the last one is the probability map computed from all training data

where $\mathbf{x} = (x_1, x_2, \dots, x_N)$ with x_i representing a continuous random variable for the i -th pixel that indicates the likelihood to be a pedestrian region, G is a N -dimensional Gaussian function, D^s is a diagonal covariance matrix to be determined later, and c is a normalization factor.

3 Pedestrian Segmentation

Given an image I , each pixel $v_i \in V$ is assigned to a label $l_i \in \{0, 1\}$ representing background and pedestrian, respectively, in the pedestrian segmentation problem. The pedestrian segmentation problem is simply to assign v_i the label based on the posteriori probability $P(x_i > \alpha | I)$, where α is set to 0.5 in this problem. Note that the shape prior $p(\mathbf{x})$ is assumed to be a mixture of Gaussian functions given in eq. 1. We will formulate the pixel likelihood estimation problem by using a graphical model framework, and derive the energy function for the image segmentation.

3.1 Graphical Model

We first describe the notion of a graph for an image. A graph $G = (V, E)$ has vertices (nodes), represented by set V , with each vertex corresponding to a pixel and $V = \{v_i\}_{i=1, \dots, N}$, and edges, denoted by $e \in E \subseteq V \times V$. An edge e connecting two vertices v_i and v_j is denoted by e_{ij} . A weighted graph has a value assigned to each edge, and it is called a weight. The weight of edge e_{ij} , is denoted by $w(e_{ij})$ and we express it as w_{ij} . The degree of a vertex is $d_i = \sum w_{ij}$ for all edges e_{ij} incident on v_i . In this work, we assume that the graph is undirected ($w_{ij} = w_{ji}$).

3.2 Edge Weights

In order to represent the image structure, one must define a function that maps a change in image intensities to edge weights. This is a common feature of graph-based algorithms for image analysis. In this work, we implement the typical Gaussian weighting function given below. Let N_i be the neighborhood of a pixel v_i . In this paper, we employ an 8-connected neighborhood structure,

$$w_{ij} = \begin{cases} \exp(-\beta(g_i - g_j)^2) & v_j \in N_i, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where g_i indicates the image intensity at pixel v_i and β is a constant that controls the strength of the weight. This equation could be modified to handle color vector data by replacing $(g_i - g_j)^2$ by $\|g_i - g_j\|^2$. The definition of weighting function is the same as random walks [1] and [10], and it provides a numerical measure for the label similarity between two neighboring pixels.

Besides the above Gaussian weighting function is used here, we also present a weighting function for the shape prior,

$$w_{ij}^s = \begin{cases} \exp(-\theta(\mu_i^s - \mu_j^s)^2) & v_j \in N_i, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

where μ_i^s denotes the probability at pixel v_i for the s-th shape prior model, and θ is a free parameter. It is similar to eq. 2. The higher value of the weight is computed when the probability of μ_i^s is very close to μ_j^s , which means that they should be labeled with the same class, and vice versa. The degree of the s-th shape prior model is $d_i^s = \sum w_{ij}^s$ for all edges e_{ij} incident on v_i . Finally, the importance for each node, denoted by D_i^s , is determined by multiplying the associated degree of the prior shape model and the degree of the same node determined by its weight, i.e.

$$D_i^s = d_i \times d_i^s \tag{4}$$

Thus, the diagonal covariance matrix D^s for the s-th Gaussian in eq. 1 is formed from the diagonal entries D_i^s .

3.3 Likelihood Estimation

Given a weighted graph, there are a set of marked nodes V_M (seeds), and a set of unmarked nodes V_U , such that $V_U \cup V_M = V$ and $V_U \cap V_M = \phi$. Therefore, we would like to label each node $v_i \in V_U$ with a class, pedestrian or background. The set of seeds V_M in previous random walker segmentation is usually obtained by user interaction. In this work, the seeds are placed automatically by using the prior human shape model, and more details will be described subsequently.

In the random walks segmentation approach, the problem is to assign a label to each node $v_i \in V_U$, the likelihood x_i , such that a random walker starting from that node first reaches a seed $v_j \in V_M$, and assigns a label for v_j . The segmentation is then completed by assigning each free node to the label corresponding to the highest likelihood. In our implementation, the node $v_i \in V_U$ will be labeled to 1 (pedestrian) if $x_i > 0.5$.

Now we review the quadratic energy function to be minimized in random walks segmentation [1],

$$E_{RW}(\mathbf{x}) = \sum_{e_{ij} \in E} w_{ij} (x_i - x_j)^2 \quad (5)$$

This energy function is called Dirichlet integral [11] in random walks. It is similar to an electrical problem, which includes three fundamental equations of circuit theory (Kirchhoff's current and voltage law and Ohm's law). Here is the explanation for this energy function in likelihood: the energy function will be minimum if the likelihood x_i and x_j at node v_i and v_j are very close when w_{ij} is a large value, thus the node v_i and v_j should be labeled to the same class.

In addition to the random walks energy function given in eq. 5, we propose another energy function to incorporate the prior shape model. The nodewise priors μ_i^s , which represents the probability of the s -th pedestrian prior model at node v_i , and the energy function of the s -th pedestrian prior model can be written as:

$$E_{Prior}^s(\mathbf{x}^s) = \sum_{v_i \in V} D_i^s (x_i^s - \mu_i^s)^2 \quad (6)$$

where x_i^s is the likelihood at pixel v_i of the s -th shape prior model.

The above two energy functions, given in eq. 5 and 6, are combined to approximate the total energy function corresponding to the MAP estimation of \mathbf{x} with the introduction of a parameter λ that controls the weighting between the two energy functions; i.e.

$$E_{Total}(\mathbf{x}) \approx E_{RW}(\mathbf{x}) + \lambda \min_{0 \leq s \leq 7} E_{Prior}^s(\mathbf{x}) \quad (7)$$

where the first term E_{RW} is the label-continuity constraint borrowed from the original Random Walks formulation enforcing that two neighboring pixels in the small neighborhood system should have the same label if their colors or intensities are similar, and the energy E_{Prior}^s is the unary constraint that each pixel tends to the s -th prior model. The weighting parameter λ is a positive coefficient measuring how much we want to fit the prior models. If $\lambda = 0$, the energy function E_{Total} is completely the same as that used in the random walks segmentation algorithm.

3.4 Convex Optimization

There is no closed-form solution to directly minimize the energy function in eq. 7. Instead, we minimize the individual combined energy function $E_{Total}^s = E_{RW} + \lambda E_{Prior}^s$ for each individual shape prior model, and find the one with the minimal cost to be the solution. For each individual combined energy function, it can be formulated as a quadratic form of \mathbf{x} as follows:

$$\begin{aligned} E_{Total}^s(\mathbf{x}^s) &= E_{RW}(\mathbf{x}^s) + \lambda E_{Prior}^s(\mathbf{x}^s) \\ &= \sum_{e_{ij} \in E} w_{ij} (x_i^s - x_j^s)^2 + \lambda \sum_{v_i \in V} D_i^s (x_i^s - \mu_i^s)^2 \\ &= \mathbf{x}^{sT} L \mathbf{x}^s + (\mathbf{x}^s - \boldsymbol{\mu}^s)^T \lambda D^s (\mathbf{x}^s - \boldsymbol{\mu}^s) \end{aligned} \quad (8)$$

where D^s is a diagonal matrix with the values D_i^s on the diagonal, $D^s = \text{diag}([D_1^s, \dots, D_N^s])$, and L represents the combinatorial Laplacian matrix [12] defined by

$$L_{ij} = \begin{cases} d_i & \text{if } i = j, \\ -w_{ij} & \text{if } v_i \text{ and } v_j \text{ are adjacent nodes,} \\ 0 & \text{otherwise,} \end{cases} \quad (9)$$

where L_{ij} is indexed by vertices v_i and v_j .

Partitioning the vertices into two sets, marked node set V_M (seeds) and unmarked node set V_U , we may decompose eq. 8 into

$$\begin{aligned} E_{Total}^s(\mathbf{x}^s) &= \mathbf{x}^{sT} L \mathbf{x}^s + (\mathbf{x}^s - \boldsymbol{\mu}^s)^T \lambda D^s (\mathbf{x}^s - \boldsymbol{\mu}^s) \\ &= \mathbf{x}_M^{sT} L_M \mathbf{x}_M^s + 2\mathbf{x}_U^{sT} B^T \mathbf{x}_M^s + \mathbf{x}_U^{sT} L_U \mathbf{x}_U^s \\ &\quad + (\mathbf{x}_M^s - \boldsymbol{\mu}_M^s)^T \lambda D_M^s (\mathbf{x}_M^s - \boldsymbol{\mu}_M^s) \\ &\quad + (\mathbf{x}_U^s - \boldsymbol{\mu}_U^s)^T \lambda D_U^s (\mathbf{x}_U^s - \boldsymbol{\mu}_U^s) \end{aligned} \quad (10)$$

where $\mathbf{x}^s = [\mathbf{x}_M^s, \mathbf{x}_U^s]$ and $\boldsymbol{\mu}^s = [\boldsymbol{\mu}_M^s, \boldsymbol{\mu}_U^s]$ correspond to the partitioning of the labels and potentials into the seeded and unseeded nodes, respectively.

Differentiating the above matrix form for the energy function E_{Total}^s given in eq. 10 with respect to \mathbf{x}_U^s , and setting it to zero yields

$$\frac{\partial E_{Total}^s}{\partial \mathbf{x}_U^s} = B^T \mathbf{x}_M^s + L_U \mathbf{x}_U^s + \lambda D_U^s (\mathbf{x}_U^s - \boldsymbol{\mu}_U^s) = 0 \quad (11)$$

then the system of linear equations can be written as

$$(L_U + \lambda D_U^s) \mathbf{x}_U^s = -B^T \mathbf{x}_M^s + \lambda D_U^s \boldsymbol{\mu}_U^s \quad (12)$$

Since the matrix $A = L_U + \lambda D_U^s$ is positive definite, the linear system in 12 can be solved easily by an iterative numerical algorithm, such as conjugate gradient, to obtain the likelihood \mathbf{x}_U^s for all unmarked pixels.

3.5 Prior Model Decision

The score function is presented for deciding the prior model automatically. The correlation coefficient is used as a score function $Score(\mathbf{x}^s, \boldsymbol{\mu}^s)$ that evaluates the normalized correlation between the prior model $\boldsymbol{\mu}^s$ and its corresponding segmentation result \mathbf{x}^s . Let s^* denote the prior model with the maximum score, i.e.

$$s^* = \text{argmax}_s (Score(\mathbf{x}^s, \boldsymbol{\mu}^s)) \quad (13)$$

Finally, after we compute the likelihood $x_i^{s^*}$ in eq. 12 for the s^* -th prior model, the decision rule of each pixel v_i for image segmentation is given as follows

$$l_{v_i}^{s^*} = \begin{cases} 1(\text{pedestrian}) & \text{if } x_i^{s^*} > T_1, \\ 0(\text{background}) & \text{otherwise,} \end{cases} \quad (14)$$

where T_1 is the threshold. After assigning the label $l_{v_i}^{s*}$ to each pixel v_i , the pedestrian segmentation is accomplished.

3.6 Human Detection and Refinement

In order to make use of the prior model appropriately, it is necessary to find the precise location of pedestrian in an image. First of all, a human detector is applied to obtain a rough position, I_r , of pedestrian in this system. Fig. 4 (a) shows the result of human detection in [13], the bounding box is regarded as a coarse pedestrian location in this procedure.

Refining the precise location of pedestrian based on the rough position is proposed in this section. This problem is formulated as a binary sliding-window search problem. A multiple-size window is scanned over the bounding box which is determined from the human detection result, and template matching provides estimates of the pose model parameters for every detected window. We define a score function, which is the same as that given in eq. 14, to find precise location of pedestrian based on the maximal score.

The random walks segmentation without prior model is applied to I_r to obtain the ordinary probability of pedestrian (Fig. 4 (c)), and the seeds are placed using $\mu_i^0 \in \boldsymbol{\mu}^0$ (prior model with whole training data), i.e.

$$Seed_{ped} = \{v_i | \mu_i^0 > T_2\} \quad (15)$$

where T_2 is threshold for shape prior, and $Seed_{background}$ is determined with respect to the rectangular boundary of I_r . An example of human detection refinement is depicted in Fig. 4 (b).

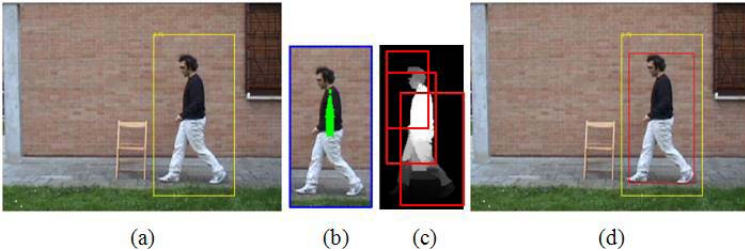


Fig. 4. Example of human detection and refinement procedure. (a) Result of human detection, the bounding box with yellow line shows I_r . (b) Seeds are placed in I_r . The green and blue dots represent the pedestrian and background seeds, respectively. (c) Result of the random walks segmentation with the seeds from (b), and the bounding boxes with red lines indicate the multiscale sliding window. (d) The final result with human detection and refinement procedure.

4 Experimental Results

All of our experiments were performed on a PC equipped with Intel i5 CPU 750 (2.67 GHz) and 2 GB memory. The proposed pedestrian segmentation algorithm

was implemented in MATLAB. We have used the MIT pedestrian dataset [14], INRIA person dataset [15], and ViSOR surveillance video dataset [16] to evaluate the results of our algorithm. The ViSOR dataset consists of surveillance videos, and we trimmed a short pedestrian sequence from it for the testing.

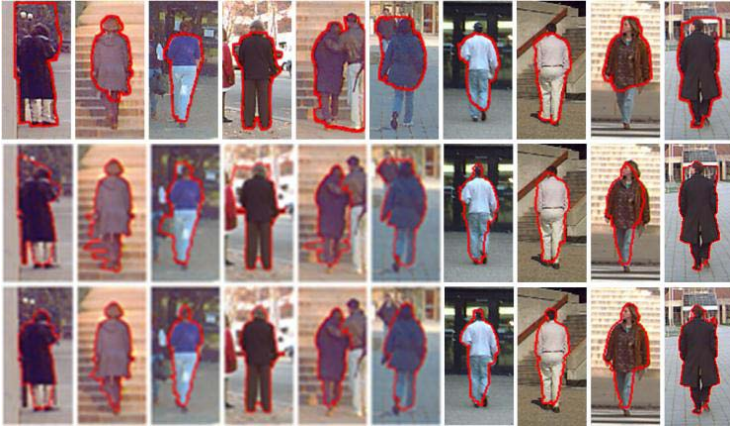
Refining the precise location of pedestrian is formulated as a sliding-window search problem. We refine the detection window by using the NCC score with 7 scales of the sliding windows. In order to quantify the results of the pedestrian segmentation, we applied our algorithm on MIT, INRIA, and ViSOR datasets, and compare the segmentation results with manually labeled masks, which are regarded as the ground truth.

The percentage of overlap area between the ground truth mask and the segmented region is evaluated as the segmentation accuracy in this experiment. The manually labeled segmentation ground truths are sometimes ambiguous around the boundary pixels. Hence we mark a two-pixel width do-not-care (DNC) boundary for accuracy assessment of the human segmentation results. This strategy is similar to that used in [18] [19].

We select 60 testing images from MIT and INRIA datasets, and cut a short video from the "Man with a dog" video sequence in the ViSOR dataset, the trimmed video contains 30 testing images, Table 1 shows comparison of the segmentation accuracies for the GrabCut [20] algorithm, the random walks algorithm and the proposed algorithm. Fig. 6 depicts some segmentation results from the testing images for the comparison of the proposed pedestrian segmentation algorithm with the GrabCut [20] algorithm and the random walks segmentation algorithm. In our experiments, all of the three algorithms have with the same initialized windows of detected humans for a fair comparison. It is obvious from Table 1 and Fig. 6 that the proposed pedestrian segmentation algorithm significantly outperforms the other two well-known segmentation algorithms in our experiments.



Fig. 5. The testing image samples and the corresponding segmentation ground truth masks



(a)



(b)



(c)

Fig. 6. Pedestrian segmentation results by using three different methods with the same initializations on testing images in (a) MIT, (b) INRIA, and (c) ViSOR datasets. The first rows give the results by using the GrabCut method, the second rows are the results by using the random walks algorithm, and the third rows show the results by using the proposed pedestrian segmentation algorithm.

Table 1. Overlap area percentages between ground truth and segmented regions

Dataset	Methods	Accuracy
MIT	GrabCut	68.06 %
	Random Walks	81.16 %
	Our method	84.36 %
INRIA	GrabCut	46.56 %
	Random Walks	69.78 %
	Our method	83.36 %
ViSOR	GrabCut	46.14 %
	Random Walks	71.03 %
	Our method	87.08 %

5 Conclusion

In this paper, we presented an automatic pedestrian segmentation algorithm by incorporating pedestrian shape prior into random walks segmentation. Our experimental results show that the proposed algorithm can provide good segmentation results for the cases with slight occlusion, similar background and illumination changes. In addition, a new pedestrian dataset with labeled silhouettes is produced.

For the directions of future research, the over-segmentation can be incorporated into this framework to obtain more accurate segmentation and reduce the computational time by taking each region as a node. It could be a possible direction for developing a more accurate and robust pedestrian segmentation technique.

References

1. Grady, L.: Random Walks for Image Segmentation Journal. *IEEE Trans. Pattern Analysis and Machine Intelligence* 28(11), 1768–1783 (2006)
2. Juan, C.-F., Chang, C.-M., Wu, J.-R., Lee, D.: Computer Vision-Based Human Body Segmentation and Posture Estimation. *IEEE Trans. Systems, Man and Cybernetics, Part A: Systems and Humans* 39(1), 119–133 (2009)
3. Cucchiara, R., Grana, C., Prati, A., Vezzani, R.: Probabilistic posture classification for Human-behavior analysis. *IEEE Trans. Systems, Man and Cybernetics, Part A: Systems and Humans* 35(1), 42–54 (2005)
4. Lin, Z., Davis, L.S.: Shape-Based Human Detection and Segmentation via Hierarchical Part-Template Matching. *IEEE Trans. Pattern Analysis and Machine Intelligence* 32(4), 604–618 (2010)
5. Lin, Z., Davis, L.S., Doermann, D., DeMenthon, D.: Hierarchical Part-Template Matching for Human Detection and Segmentation. In: *International Conf. on Computer Vision*, pp. 1–8 (2007)
6. Gao, W., Ai, H., Lao, S.: Adaptive Contour Features in oriented granular space for human detection and segmentation. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 1786–1793 (2009)

7. Viola, P., Jones, M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: International Conf. on Computer Vision, vol. 2, pp. 734–741 (2003)
8. Schapire, R., Singer, Y.: Improved Boosting Algorithms Using Confidence-rated Predictions. *Machine Learning* 37, 297–336 (1999)
9. Givoni, I.E., Frey, B.J.: A Binary Variable Model for Affinity Propagation. *Neural Computation* 21, 1589–1600 (2009)
10. Kim, T.H., Lee, K.M., Lee, S.U.: Nonparametric higher-order learning for interactive segmentation. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 32201–3208 (2010)
11. Courant, R., Hilbert, D.: *Methods of Math. Physics*, vol. 2. John Wiley and Sons (1989)
12. Merris, R.: Laplacian Matrices of Graphs: A Survey. *Linear Algebra and Its Applications* 197,198, 143–176 (1994)
13. Maji, S., Berg, A.C., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
14. <http://cbcl.mit.edu/cbcl/software-datasets/PedestrianData.html>
15. howpublished, <http://pascal.inrialpes.fr/data/human/>
16. howpublished, <http://www.openvisor.org/>
17. <http://iris.usc.edu/Vision-Users/OldUsers/bowu/DatasetWebpage/dataset.html>
18. Shotton, J., Winn, J.M., Rother, C., Criminisi, A.: TextonBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) *ECCV 2006*. LNCS, vol. 3951, pp. 1–15. Springer, Heidelberg (2006)
19. Wu, B., Nevatia, R.: Simultaneous Object Detection and Segmentation by Boosting Local Shape Feature based Classifier. In: IEEE Conf. on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
20. Rother, C., Kolmogorov, V., Blake, A.: Interactive Foreground Extraction using Iterated Graph Cuts. *ACM Trans. on Graphics* 23, 309–314 (2004)
21. Grady, L.: Multilabel Random Walker Image Segmentation Using Prior Models. In: IEEE Conf. on Computer Vision and Pattern Recognition, vol. 1, pp. 763–770 (2005)

A Novel Rate Control Algorithm for H.264/AVC Based on Human Visual System

Jiangying Zhu, Mei Yu, Qiaoyan Zheng, Zongju Peng,
Feng Shao, Fucui Li, and Gangyi Jiang

Faculty of Information Science and Engineering, Ningbo University
315211 Ningbo, China

Abstract. To improve performance of rate control algorithm for H.264/AVC, and keep a better control accuracy of the output of the compressed video stream, a novel rate control algorithm based on human visual system (HVS) is proposed in this paper. The proposed rate control algorithm consists of two layers: frame level and basic unit (BU) level. In frame level, changed scene is first detected and frame difference ratio is utilized to represent the motion complexity of the frame, and then target bit for frame level are allocated by considering the two factors. In BU level, by analyzing motion information, texture characteristics and the location of the frames, visual sensitivity of a macroblock is first measured, and the bit is allocated for the macroblock based on the sensitivity factor. Experimental results show that the proposed method can provide an improved visual quality and higher PSNR while almost the same control accuracy, compared with traditional rate control method.

Keywords: H.264/AVC, rate control, scene change detection, visual sensitivity.

1 Introduction

In recent years, it is crucial to maintain a good balance between the allocated rate and the video quality under constraint of bandwidth requirement. Rate control has been widely studied in digital video coding standards and applications, such as TM5 for MPEG-2 [1], TMN8 for H.263 [2], VM8 for MPEG-4 [3] and JVT-G012 for H.264/AVC [4]. It dynamically adjusts encoder parameters to achieve a target bit rate and effective rate control algorithm can result in high video quality, low video quality fluctuation, and a low mismatch between the target and the actual encoded bit rates.

In the literatures, many RC algorithms had been proposed to improve the quality in H.264 coding. The JVT-G012 model used a fluid flow traffic model to compute the target bits for the current encoding frame and a linear model to predict mean absolute difference (MAD) to solve the chicken and egg dilemma. To improve performance of rate control algorithm for JVT-G012, Lee et al. used the structural and statistical features of local textures to adaptively set proper initial QP values for versatile video contents and peak signal-to-noise ratio (PSNR) variation-limited bit-allocation [5]. In addition, Sun et al. proposed a new rate-complexity-quantization model and an

incremental rate control algorithm for H.264/AVC video coding [6], by estimating the picture complexity and rate-quantization modeling in an incremental rate control for P-frames. Xie *et al.* used texture-complexity as an innovative coding characteristic [7], based on the discrete cosine transform (DCT) coefficients distribution and distortion-quantization (DQ) relationship, to propose a general rate-distortion (RD) model for block-based video coding. There are many works aimed at reducing computational complexity and improving prediction accuracy [8-10]. Zhao *et al.* presented an effective macro block layer rate control algorithm [11], analyzing the temporal-spatial correlation and object direction, and calculating QP based on a more accurate header bits prediction model. However, these methods do not take into account human visual system in rate control; therefore, many scholars recently carry out deeply research in this connection. Chen *et al.* proposed a perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion [12]. Tang *et al.* used the loss of human visual sensitivity model to allocation the bits [13]. The common approach is based on region of interest rate control algorithm [14]. Zheng *et al.* proposed a human visual system based rate control algorithm [15], by analyzing the motion complexity of the frames and human visual sensitivity, to control bit allocation. For some sequences with high motion or scene changes, these approaches failed to predict the abrupt changes of MAD. In order to improve these problems, we propose a more accurate MAD model based on human visual system (HVS).

In this paper, we propose a simple and effective model based on HVS, especially for high motions areas or scene changes areas. The frame difference ratio between the current frame and its previous frame is defined to describe frame complexity, while visual sensitivity factor is defined to describe BU complexity, which based on motion information, texture characteristics and the location of the frames. Meanwhile, by thoroughly analyzing the temporal-spatial correlation, a more accurate MAD model is presented to improve the quality of frames. The experimental results show the effectiveness of the proposed scheme.

2 Analysis of the Rate Control Algorithm in JVT-G012

There are three layers in H.264/AVC bits allocation: GOP-layer, frame-layer and BU layer. The rate control algorithm in JVT-G012 is adaptive BU layer rate control. BU can be a frame, a slice, or an MB. The algorithm is described as follows.

1) Compute the target bit for the current frame

The target bits allocated for each frame in the GOP is determined based on buffer occupancy and the number of remaining bits. Meanwhile, in order to keep consistent the visual quality, the relative complexity of the current frame also is considered. T_i is the number of target bits generate by the i th GOP, which is computed by

$$T_i = \beta \times T_i' + (1 - \beta) \times T_i'' \quad (1)$$

where β is a constant and its typical value is 0.5 if there is no B frame, T_i' is the number of frame bits which is computed by using the hypothetical reference decode

(HRD) model, and T_i'' is the number of frame bits which is computed by remaining bits in the video sequence and the relative complexity in the frame (W_p, W_b), N_{pr} and N_{br} are the number of remaining bits of P frame and B frame, which is computed by

$$T_i'' = \frac{W_p T_r}{W_p N_{pr} + W_b N_{br}} \quad (2)$$

In the case that there is no B frame, equation (2) can be simplified as

$$T_i'' = \frac{T_r}{N_{pr}} \quad (3)$$

2) Allocate the target bit for the basic unit

In JVT-G012, a linear model is used to predict the MAD of current basic unit in the current frame by that of the co-located basic unit of the previous frame. Suppose that the predicted MAD of current basic unit in the current frame and the actual MAD of the co-located basic unit in the previous frame are denoted by MAD_{cb} and MAD_{pb} , respectively. The linear prediction model is then given by

$$MAD_{cb} = a_1 \times MAD_{pb} + a_2 \quad (4)$$

where a_1 and a_2 are two coefficients of the prediction model. The initial values of a_1 and a_2 are set to 1 and 0, respectively. They are updated after coding each basic unit. If T_{f_r} is the number of remaining bits of previous frame, N is the number of BU in the current frame. T_{f_i} is the target bit for the current BU is computed by

$$T_{f_i} = T_{f_r} \times \frac{MAD_i^2}{\sum_{j=i}^N MAD_j^2} \quad (5)$$

3) Compute the corresponding parameter by using the quadratic R-D model

The parameter of the used quadratic R-D model is computed by

$$T - H = \left(\frac{C_1}{Q_{step}} + \frac{C_2}{Q_{step}^2} \right) \times MAD \quad (6)$$

where T is the total number of target bits, H is the header bits, C_1 and C_2 are two coefficients of the model, Q_{step} is quantization step.

If there is no B frame in equation (3), the remaining bits are allocated to all frames equally, while ignoring the image complexity. Therefore, there is a bigger difference if the sequences with high motion or scene changes, resulting in significant fluctuations with the PSNR, especially at low bit rate.

In equation (4), there are two problems for MAD prediction. One is the reference MAD information used in the model, only co-located MB in the previous frame selected to predict the MAD value. For some sequences with high motion or scene

changes, this prediction model will fail to predict the abrupt changes of MAD, consequently resulting in a weak rate control. The other is the two coefficients of prediction model updated by linear regress will introduce the computation cost to the already existed computation-demanding H.264/AVC encoder.

3 Improved Rate Control Method Based on HVS

3.1 Bits Allocation at Frame Level

To better allocate the target bits for every frame video that consider scene change or high motion and the complexity of current frame content. More bits is allocate to scene change frames or high complexity frames, and fewer bits for low complexity frames or unimportance frames to achieve high video quality.

There are some types of video scene changes: mutation scene change, melting, fading and so on. There are three algorithms with fast scene change detection: the detection based on gray value, motion and search for edges. The latter two kinds of detection algorithms with good performance, but the algorithm of high complexity greatly limits their applications, especially in the high demand for real-time video communication rate control algorithm. Accordingly, the current frame and reference frame use the three components of the mean absolute difference in the current frames for judging whether a scene change, the difference function is given by

$$D_x(S_{cur}, S_{ref}) = |mean(S_{cur}, x) - mean(S_{ref}, x)| \quad (7)$$

where $mean(X)$ denotes mean function, S_{cur} and S_{ref} are the reconstructed frames of the current frame and previous frame; X is the component in YUV. To determine whether a scene change, equations (8) and (9) are used.

$$\frac{D_Y(S_i, S_{i-1})}{mean(S_i, Y)} + \frac{D_U(S_i, S_{i-1})}{mean(S_i, U)} + \frac{D_V(S_i, S_{i-1})}{mean(S_i, V)} \geq t_{TH1} \quad (8)$$

$$D_Y(S_i, S_{i-1}) + D_U(S_i, S_{i-1}) + D_V(S_i, S_{i-1}) \geq t_{TH2} \quad (9)$$

The current frame is scene change frame when equations (8) and (9) are both true, t_{TH1} and t_{TH2} are the decision threshold, which are computed by

$$t_{TH1} = \frac{|mean(D_Y(S_i, S_{i-1}))|}{mean(Y)} + \frac{|mean(D_U(S_i, S_{i-1}))|}{mean(U)} + \frac{|mean(D_V(S_i, S_{i-1}))|}{mean(V)} \quad (10)$$

$$t_{TH2} = mean(D_Y(S_i, S_{i-1})) + mean(D_U(S_i, S_{i-1})) + mean(D_V(S_i, S_{i-1})) \quad (11)$$

We use frame difference function describe the frame complexity, shown as

$$D_i = \sum_{y=0}^{y=H-1} \sum_{x=0}^{x=W-1} |I_C(x, y) - I_P(x, y)| \tag{12}$$

where $I_c(x, y)$ and $I_p(x, y)$ are the luminance value of the current frame and previous frame in the pixel (x, y) , D_i is the i th frame of the frame difference. In general, if there is strong motion or scene change between two frames, there is larger value of the frame difference. Then, Eq.(3) is further bounded

$$T_i^n = \alpha \times \frac{T_r}{N_{p,r}} \tag{13}$$

$$\alpha = \min \left\{ \max \left\{ \frac{D_i}{D_{i-1}}, k_1 \right\}, k_2 \right\} \tag{14}$$

where a is the frame complexity, D_i is the frame difference value of the i th frame that is non-coded, and D_{i-1} is the frame difference value of previous frame. The value of coefficient a has a limit region; the constants k_1 and k_2 come from experiments. Separately, if the current frame detected as scene change, $a = \text{MAX}(k_1, k_2)$. To smooth the visual quality, the value of k_1 and k_2 are typical set to 0.5 and 2.

3.2 Bits Allocation at BU Level

To achieve good subject quality, rate control should consider the characters of Human Visual System. As we all known, the regions of interest often focus on the middle position of the image, the movement of objects and complex texture of the object, when we note a video sequence. The same amount of MAD in the video sequence of different regions will have distinct subjective experience: visually sensitive areas of the visual distortion caused significantly greater than in other regions. Therefore, taking into account the human visual systems, in this paper we first define a reasonable visual factor in the BU level for target bit allocation.

1) Visual sensitivity measure model

Visual sensitivity of each MB is corresponded with position, movement and the complexity of the texture. In order to get visual weight conform to human visual sensitivity, if it is sensitivity to human, the value of weight great than 1, otherwise less than 1. In the JVT-G012, the visual weight in BU layer are the same (namely the weighting factor W is set to 1). Firstly, we should compute the value of W , $W \in [W_{\text{MIN}}, W_{\text{MAX}}]$ ($W_{\text{MIN}} < 1$, $W_{\text{MAX}} > 1$), then, using W to allocate target bits at BU layer. While BU is composed by several macro blocks, we should calculate the location of each macro block, and the complexity of texture.

Gaussian function is used to determine the location weight of each MB, due to the objects in the central region of the image more attractive eyes than in its external edge

of the area. Suppose that the coordinate of the image center is (x_c, y_c) , MB's location weight at (x, y) is defined by

$$L_{MB}(x, y) = \exp\left(-\frac{(x - x_c)^2 + (y - y_c)^2}{2\sigma^2}\right) \tag{15}$$

where σ is the scale parameter of Gaussian function, the smaller the value of σ , the faster the speed of reducing. If $\delta = \min((W-1)/2, (H-1)/2)$, equation (15) exactly describes the location weight of the macro block, where W and H are image's weight and height.

Here, we use mean absolute difference between in the co-located position of the current and previous frames to describe motion degree of MB, in order to measure the activity of the MB, which is shown in Eq. (16)

$$A_{MB}(n, m) = \frac{1}{256} \sum_{i,j}^{16} |f_{n,m}(i, j) - f_{n-1,m}(i, j)| \tag{16}$$

where $f_{n,m}(i, j)$ and $f_{n-1,m}(i, j)$ denote the luminance value in the pixel (i, j) of the m th MB in the n th and $(n-1)$ th frame. Obviously, the movement area with a large value and the method will not bring too much computational complexity. Therefore, the MB activities measure has a new definition with the location weight of each MB (L_{MB})

$$A_{MB}' = L_{MB} \times A_{MB} \tag{17}$$

Then $mot(i, j)$ denotes regular of the MB activities, as shown in Eq. (18)

$$mot(i, j) = \frac{2 \times A(i, j) + A_{avg}}{A(i, j) + 2 \times A_{avg}} \tag{18}$$

where $A(i, j)$ is the activities of the j th BU in the i th frame, A_{avg} is the average activities of the remaining BU in current frame.

Due to take into account the human eyes are more sensitive to the distortion on the edges of objects, we use gradient to get the object edges as the complexity of the MB texture. While the complex dynamic objects in the static area are not attractive eye's attention like in the dynamic region, so we directly detect edge to the difference image Δf between adjacent frames rather than the original image, and then the difference for each pixel separately do convolution with two operators, it will create a corresponding edge vector $\overline{Grad}(i, j) = (gradx(i, j), grady(i, j))$, which are computed by

$$gradx(i, j) = \Delta f(i - 1, j + 1) + 2\Delta f(i, j + 1) + \Delta f(i + 1, j + 1) - \Delta f(i - 1, j) - 2\Delta f(i, j - 1) - \Delta f(i + 1, j - 1) \tag{19}$$

$$grady(i, j) = \Delta f(i + 1, j - 1) + 2\Delta f(i + 1, j) + \Delta f(i + 1, j + 1) - \Delta f(i - 1, j - 1) - 2\Delta f(i - 1, j) - \Delta f(i - 1, j + 1) \tag{20}$$

where $\Delta f(i, j)$ denotes the luminance value of the pixel (i, j) in the difference image Δf . So the edge vector of the pixel (i, j) in the difference image Δf is computed by

$$Grad(i, j) = \sqrt{gradx(i, j)^2 + grady(i, j)^2} \quad (21)$$

The complexity of the MB texture ($text_{MB}$) denotes the edge vector average value of all pixels in MB is computed as follows

$$text_{MB} = \frac{1}{256} \sum_{i,j}^{16} Grad(i, j) \quad (22)$$

The complexity of the MB is weighted with the location weight of each MB (L_{MB})

$$text_{MB} = L_{MB} \times text_{MB} \quad (23)$$

Then $texture(i, j)$ denotes the texture complexity of the j th BU in the i th frame

$$texture(i, j) = \frac{2 \times text(i, j) + text_{avg}}{text(i, j) + 2 \times text_{avg}} \quad (24)$$

Let $W(i, j)$ denote visual sensitivity of the j th BU in the i th frame, and it is computed by

$$W_{(i,j)} = w_1 \times mot(i, j) + w_2 \times texture(i, j) \quad (25)$$

where w_1 and w_2 are the weighting coefficients, whose typical values must fulfill the conditions: $0 \leq w_1, w_2 \leq 1$ and $w_1 + w_2 = 1$, which come from experiments.

2) A novel MAD prediction model

Based on the previous analysis, MAD predictions may be not very accurate, so a new MAD prediction model is presented by

$$MAD_{cb} = a_1 \times MAD_{pb}' + a_2 \quad (26)$$

where MAD_{pb}' denotes a new measure for evaluating the difference between the current original frame and previous reconstructed frame. In order to reduce the computational complexity, we use three kinds of MB included co-located MB, left MB and upper MB to predict MAD for current MB, which is computed by

$$MAD_{pb}' = \frac{a \times MAD_{pb} + b \times MAD_L + c \times MAD_U}{a + b + c} \quad (27)$$

where a, b and c represent the weighting coefficients of a MB, and $a+b+c=1$. The MAD_L and MAD_U are the actual MAD values of the left-side MB and upper-side MB.

If the current MB is the first MB in the current Frame, Eq. (27) is equal to

$$MAD1 = MAD_p \tag{28}$$

If the current MB is the first row MB in the current Frame, Eq. (27) is equal to

$$MAD2 = (1 - a) \times MAD_p + a \times MAD_L \tag{29}$$

If the current MB is the first column MB in the current Frame, Eq. (27) is equal to

$$MAD3 = (1 - a) \times MAD_p + a \times MAD_U \tag{30}$$

3) Improved Bits Allocation Scheme at BU level

W is a simple and accurate measure of frame complexity, therefore, it can provide a mechanism to control estimation of the target bit. If the value of W is bigger, it also means that the current BU can better attract human eyes attention, the target bits of BU is computed by

$$T_{(i,j)} = T_{f_r} \times W_{(i,j)} \times \frac{MAD_{(i,j)}^2}{\sum_{l=j}^N MAD_{(i,l)}^2} \tag{31}$$

where $T(i, j)$ denotes the total bits of the j th BU in the i th frame, T_{f_r} denotes the number of remaining bits in the current frame, $MAD(i, j)$ denotes MAD of the j th BU in the i th frame, N is the number of BU in the current frame.

Comparing the header bits with the texture bits, if the texture bits are less than header bits, $QPcb = QPpb + 2$; on the contrary, it will get QP based on R-D model of the H.264/AVC rate control.

4 Experimental Results and Analyses

To evaluate performances of the proposed method, the test platform is JM10.1 [16]; JVT-G012 is used as a benchmark for comparison. Rate control experiments are implemented on seven QCIF video sequences with different activity and motion features. The target bit rate is set to 32, 48 and 64, a, b, c in Eq.(27) is set 0.8, 0.1 and 0.1. Respectively, some important test conditions are listed in Table 1. The rate control error is used to measure the accuracy of the bitrate estimation, and defined by

$$E_{rate_predict} = \frac{|R_{acture} - R_{predict}|}{R_{predict}} \times 100\% \tag{32}$$

where R_{acture} is the bit rate of the test sequence, and $R_{predict}$ is the target bit rate.

In Table 2, compared with JVT-G012, the proposed algorithm achieves higher average PSNR and more accurate target bit rates. It is clear that JVT-G012 has a rate error range from 0.52% to 1.53%, while the proposed method ranges from 0.27% to

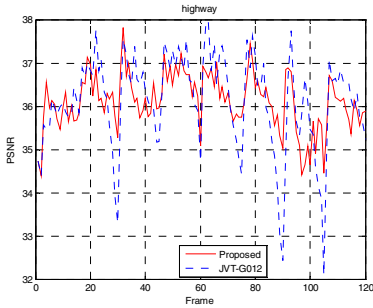
Table 1. Test conditions

Hadamard, RDO	used
Frame rate	30f/s
MV search range	16 (QCIF)
Entropy coding	CABAC
Reference frames	5
Sequence type	IPPP.....
GOP length	15
Number of frame	120

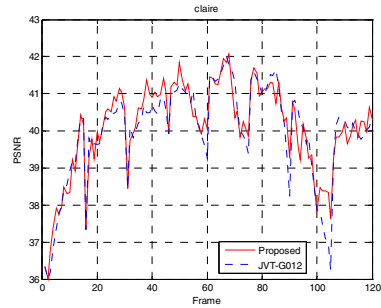
Table 2. Performance comparison between JVT-G012 and the proposed method

Target bit rate (Kbps)	Sequence	Bit rate					PSNR (dB)		
		Actual(Kbps)		Rate error (%)		Rate error gain (%)	G012	Pro.	Gain
		G012	Pro.	G012	Pro.				
32	Highway	32.33	32.13	1.03	0.41	0.62	34.39	34.72	0.33
	Claire	32.2	32.16	0.63	0.50	0.13	37.56	37.8	0.24
	Carphone	32.29	32.12	0.91	0.37	0.54	30.87	31.03	0.16
	foreman	32.24	32.13	0.75	0.41	0.34	27.89	27.97	0.08
	monitor	32.22	32.22	0.69	0.69	0.00	31.85	31.58	-0.27
	mother-daughter	32.15	32.19	0.47	0.59	-0.12	34.12	34.22	0.1
	news	32.18	32.14	0.56	0.44	0.12	29.71	29.78	0.07
48	Highway	48.63	48.29	1.31	0.60	0.71	36.11	36.11	0
	Claire	48.39	48.19	0.81	0.40	0.41	39.99	40.09	0.1
	Carphone	48.44	48.15	0.92	0.31	0.61	32.97	33.19	0.22
	foreman	48.27	48.19	0.56	0.40	0.16	30.46	30.39	-0.07
	monitor	48.38	48.29	0.79	0.60	0.19	34.16	34.26	0.1
	mother-daughter	48.5	48.23	1.04	0.48	0.56	35.87	36.03	0.16
	news	48.31	48.15	0.65	0.31	0.34	31.9	31.98	0.08
64	Highway	64.98	64.23	1.53	0.36	1.17	37.03	37.07	0.04
	Claire	64.34	64.25	0.53	0.39	0.14	41.64	41.74	0.1
	Carphone	64.55	64.21	0.86	0.33	0.53	34.43	34.71	0.28
	foreman	64.56	64.27	0.88	0.42	0.46	32.07	32.05	-0.02
	monitor	64.37	64.31	0.58	0.48	0.10	35.92	36	0.08
	mother-daughter	64.59	64.36	0.92	0.56	0.36	37.35	37.43	0.08
	news	64.33	64.17	0.52	0.27	0.25	33.5	33.75	0.25

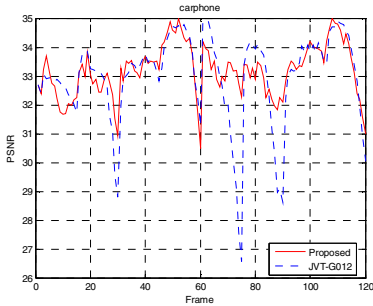
0.69%, the maximum of rate error gain is 1.17%. PSNR of the proposed method is improved up to 0.33dB compared with JVT-G012. In these sequences, the sequence that scene changes intense is Highway sequence, resulting in the most of the error rate decreased, smaller PSNR fluctuation, because of the frame difference in the frame layer. Second is the Carphone, the Claire is the last, another reason is the use of the human visual system in the BU level control. While in the lower bit rates, the Monitor sequence, which the texture of the around regions are more complex than the middle, each BU requires a lot of bit rates. The foreman sequence which motion is more violent, each frame requires a lot of bit rates.



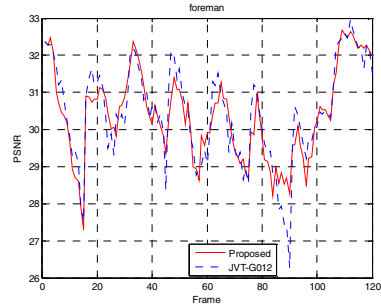
(a) Highway: Pro. =0.91, G012=1.10



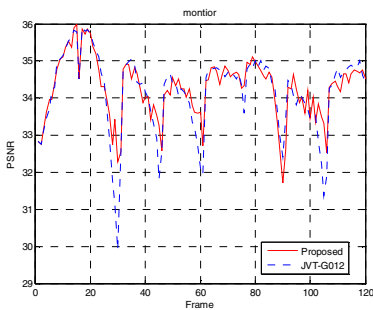
(b) Claire: Pro. =1.23, G012=1.33



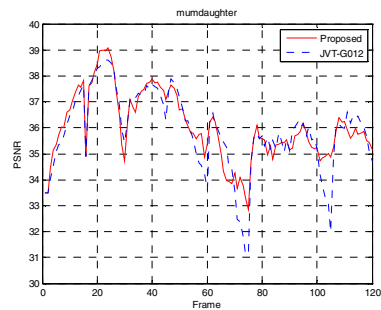
(c) car phone : Pro.=0.91, G012=1.60



(d) foreman: Pro.=1.21, G012=1.19



(e) Monitor: Pro. =0.81, G012=1.08



(f) mother-daughter : Pro.=1.33,G012=1.56

Fig. 1. PSNR fluctuation comparison between the two RC algorithms

In order to provide a specific comparison between JVT-G012 and the proposed rate control scheme, some experimental results are shown in Figs.1-3. Fig.1 illustrated that the proposed method can avoid drastic visual quality variation caused by scenes. Smaller PSNR fluctuation implies more stable visual quality which is highly desired in video coding. Figs.2-3 shows that the actual video sequences demonstrate that our algorithm suits real applications. From visual comparison, our proposed algorithm can select a more precise method to provide an acceptable visual image quality. It is clear that our proposed algorithm can achieve a higher PSNR than the existing algorithms and reduces the bit rate mismatch ratio.

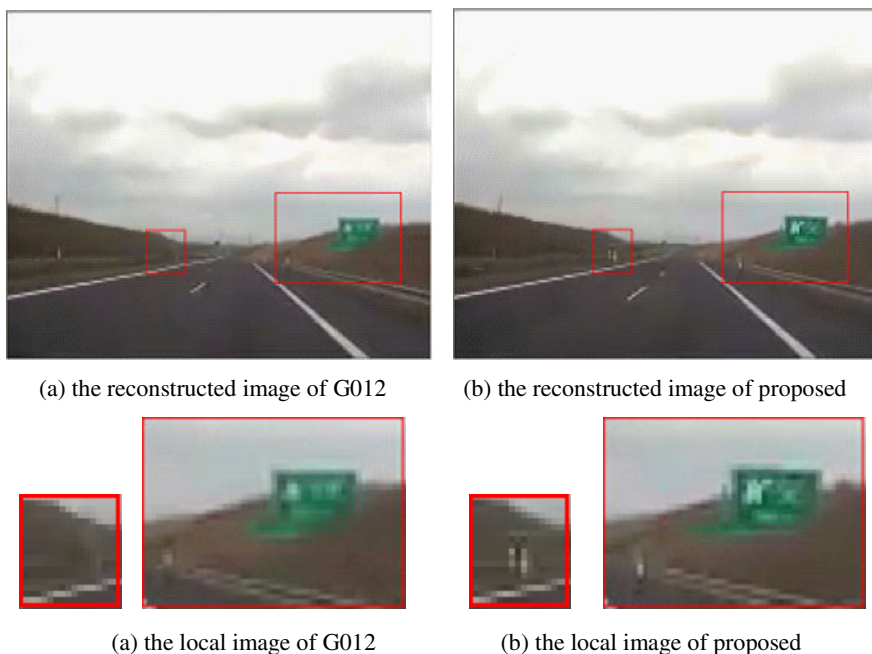


Fig. 2. Subjective visual comparison of two RC algorithms, in the 87th frame of “Highway”

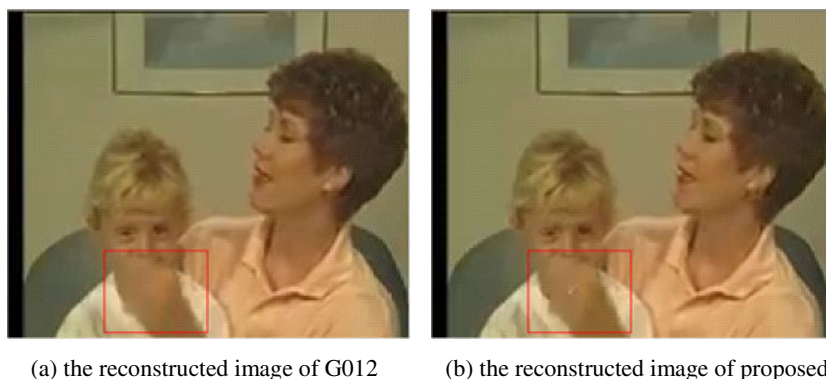


Fig. 3. Subjective visual comparison of two RC algorithms at the 59th frame of “mother-daughter”



(a) the local image of G012



(b) the local image of proposed

Fig. 3. (Continued)

5 Conclusion

This paper has presented an efficient rate control algorithm based on human visual system for a H.264/AVC video coding. By analyzing the model of JVT-G012 rate control, it is clear that scene's characteristics and its bit allocation is more reasonable so that it can maintain a video stream with a smoother PSNR variation which is highly desirable in real-time video coding and transmission. Experimental results show that compared with JVT-G012 in the low bit rate, the proposed algorithm achieves higher average PSNR and more accurate rate control.

In future work, the proposed algorithm will be extended to rate control in stereo video communication, such as mobile 3DTV. In stereo video algorithm, the bit rate budget is first allocated to each pair of stereo image frame adaptively updated according to bandwidth and buffer status, combined with human visual system of stereo image frame. Then it should also allocate bits among views according to frame complexity to keep the quality of each view. It may be helpful to get a more consistent visual quality when scene switching occurs in stereo video.

References

1. MPEG-2 Test Model 5, Doc. ISO/IEC JTC1/SC29 WG11/93-400 (April 1993)
2. Corbera, J.R., Lei, S.: Rate control in DCT video coding for low delay communication. *IEEE Transactions on Circuits and Systems for Video Technology* 9(1), 172–185 (1999)
3. Lee, H.J., Chiang, T.H., Zhang, Y.Q.: Scalable rate control for MPEG-4 video. *IEEE Transactions on Circuits and Systems for Video Technology* 10(6), 878–894 (2000)
4. Li, Z.G., Pan, F., Lim, K.P., et al.: Adaptive basic unit layer rate control for JVT. In: *The 7th JVT Meeting, JVT-G012-rl, Thailand* (2003)
5. Lee, G.G., Lin, H.Y., Wang, M.J.: Rate control algorithm based on intra-picture complexity for H.264/AVC. *IET Image Process.* 3(1), 26–39 (2009)
6. Sun, Y., Zhou, Y., Feng, Z., He, Z., Sun, S.: Incremental rate control for H.264/AVC video compression. *IET Image Processing* 3(5), 286–298 (2009)
7. Xie, Z., Bao, Z., Xu, C., Zhang, G.: Optimal bit allocation and efficient rate control for H.264/AVC based on general rate-distortion model and enhanced coding complexity measure. *IET Image Processing* 4(3), 172–183 (2010)
8. Tiago, A.F., Ricardo, L.Q.: Complexity-scalable H.264/AVC in an IPP-based video encoder. In: *IEEE International Conference on Image Processing* (2010)

9. SanzRodríguez, S., del Ama-Esteban, O., de Frutos-López, M., de María, F.D.: Cauchy-density-based basic unit layer rate controller for H.264/AVC. *IEEE Transactions on Circuits and Systems for Video Technology* 20(8), 1139–1143 (2010)
10. Chen, F.C., Hsu, Y.P.: An adaptive content based H.264/AVC rate control in low bit rate video. *International Journal of Electronics and Communications* 65(6), 516–522 (2011)
11. Zhao, D.D., Zhou, Y.J., Wang, D.Y, Mao, J.F.: Effective macro block layer rate control algorithm for H.264/AVC. *Computers and Electrical Engineering* (2011)
12. Chen, Z.Z., Guillemot, C.: Perceptually-friendly H.264/AVC video coding based on foveated just-noticeable-distortion model. *IEEE Transactions on Circuits and Systems for Video Technology* 20(6), 806–819 (2010)
13. Tang, C.W.: Spatiotemporal visual considerations for video coding. *IEEE Transactions on Multimedia* 9(2), 231–238 (2007)
14. Liu, Y., Li, Z.G., Soh, Y.C.: Region-of-interest based resource allocation for conversational video communication of H.264/AVC. *IEEE Transactions on Circuits and Systems for Video Technology* 18(1), 134–139 (2008)
15. Zheng, Q.Y., Yu, M., Peng, Z.J., Shao, F., Li, F.C., Jiang, G.Y.: Human Visual System Based Rate Control Algorithm for H.264/AVC. *Journal of Optoelectronics · Laser* 22(3) (2011)
16. JM Reference Software Version 10.1.,
<http://iphone.hhi.de/suehring/tml/download/>

Blind Image Deblurring with Modified Richardson-Lucy Deconvolution for Ringing Artifact Suppression

Hao-Liang Yang, Yen-Hao Chiao, Po-Hao Huang, and Shang-Hong Lai

Dept. of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
lai@cs.nthu.edu.tw

Abstract. In this paper, we develop a unified image deblurring framework that consists of both blur kernel estimation and non-blind image deconvolution. For blind kernel estimation, we propose a patch selection procedure and integrate it with a coarse-to-fine kernel estimation algorithm to develop a robust blur kernel estimation algorithm. For the non-blind image deconvolution, we modify the traditional Richardson-Lucy (RL) image restoration algorithm to suppress the notorious ringing artifact in the regions around strong edges. Experimental results on some real blurred images are shown to demonstrate the improved efficiency and image restoration by using the proposed algorithm.

1 Introduction

Motion blur is caused by relative motion between the camera and the scene during exposure. The real camera motion is usually too complicated to estimate from a blurred image when it involves camera rotation or large scene depth variations. To simplify the problem formulation, previous researches usually assumed the camera motion is perpendicular to the optical axes and the effect of scene depth variation can be neglected. In other words, the blur kernel is assumed to be spatially invariant. Under this assumption, a blurred image, B , can be modeled as (1), where K is the blur kernel, I is the clear image, N is the noise, and \otimes is the convolution operator.

$$B = I \otimes K + N. \quad (1)$$

The blind image restoration problem in (1) is ill-posed because I and K are highly under-constrained and there are infinitely many possible combinations of I and K such that their convolution is equal to the blurred image B . Fergus et al. [4] proposed to utilize ensemble learning to estimate the blur kernel with a sophisticated variational Bayes inference algorithm, which employs the property of specific distributions of image gradients for natural images to approximate the posterior distribution. Levin [6] also exploited image statistics for estimating blur kernels. Shan et al. [9] proposed two probabilistic models to improve image restoration. One is to model the spatially random distribution of noise, and the other is a smoothness prior model which can reduce the ringing artifacts. Cho and Lee [11] proposed a latent image prediction step, which applied shock filter to recover the sharp edge information for estimating the blur kernel.

Even with a known blur kernel, the restored image may contain some undesirable reconstruction artifacts, such as the ringing artifacts. To overcome this problem, Levin et al. [7] modeled the sparse image derivative distribution as a heavy-tailed function to alleviate the ringing artifacts. Shan et al. [9] proposed a local smoothness prior which assumes the gradients of smooth regions in a blurred image are similar to those in a clear image. Yuan et al. [8, 10] proposed the concept of residual deconvolution and modified the standard Richardson-Lucy (RL) algorithm [1], [2] by incorporating either a gain-control process [7] or a bilateral-filtering-like process [9] for suppressing the ringing artifacts.

In this paper, we propose a simple and efficient method to estimate the blur kernel under the assumption of spatially-invariant case, and modify the Richardson-Lucy algorithm to form a new deconvolution method called GARL, which effectively reduces the ringing artifact. The contributions of this paper are listed as follows:

1. We propose a patch selection scheme to choose a suitable region from the input blurred image for kernel estimation with the purpose of computational efficiency.
2. Combining several popular concepts in blind kernel estimation, we develop our method using a quadratic smoothness prior, bilateral filtering, and a good patch.
3. We exploit the gradient attenuation concept and modify the standard RL algorithm to suppress the ringing artifacts in the RL-based image deconvolution.
4. We propose an iterative detail recovery procedure that can recover missing details due to ringing suppression.

The rest of this paper is organized as follows: The blur kernel estimation algorithm is introduced in section II. Non-blind image deconvolution method is proposed in section III. Experimental results are reported in section IV. Finally, we conclude in section V.

2 Blur Kernel Estimation

Several recent researches have proposed novel and effective ways to estimate the blur kernel. We integrate some of these methods and add other new procedures to form our own kernel estimation method, of both efficiency and accuracy, as depicted in Fig. 2.

In [11], the authors minimized the objective function with a quadratic regularization term using conjugate gradient method. They also adopt bilateral filtering [3] to filter out possible noises in the latent image in their framework. Their algorithm is simple, straightforward, and efficient. In [12], a new metric to measure the usefulness of image edges in motion deblurring is proposed. They found that some regions in an image are good for kernel estimation, while some are not. As a result, they construct a map telling which parts of the image are useful. During the kernel estimation, this map is used as a mask so that only the useful parts are taken into consideration.

We take the advantages of the above methods in our kernel estimation algorithm, including quadratic objective function, bilateral filtering, and the map of useful gradients. In addition, we add a new procedure – patch selection. Patch selection has

been commonly used in deblurring to reduce the execution time in blur estimation, but few researchers have put efforts to analyze it. An image of high resolution, say, a million pixels, takes a long time to be processed, so most of the time we choose one or more small patches for the blur kernel estimation instead of using the entire image. If the patch is selected well, an accurate kernel can still be found in a relatively short time. A good patch for deblurring should contain strong edges of various directions. An edge that is parallel to the blur direction provides no information, but an edge perpendicular to the blur direction is the most appropriate one.

To automatically select the patch, we adopt the concept from the Harris corner detector [13], which determines if a pixel is a good corner from the eigenvalues of the gradient covariance matrix for a local neighborhood. For a pixel \mathbf{p} , a 2x2 Harris matrix \mathbf{C} is defined as

$$\mathbf{C}(\mathbf{p}) = \sum_{(i,j) \in W_p} \begin{bmatrix} I_x^2(i,j) & I_x(i,j)I_y(i,j) \\ I_x(i,j)I_y(i,j) & I_y^2(i,j) \end{bmatrix}, \quad (2)$$

where W_p is a local window centered at \mathbf{p} , and I_x and I_y represent the partial derivatives along x and y directions, respectively. The corner response function is then defined by

$$R = \det(\mathbf{C}) - k(\text{trace}(\mathbf{C}))^2, \quad (3)$$

where k is a constant. A high response value at pixel \mathbf{p} means that the region around \mathbf{p} is probably a corner region because it contains a set of image gradients with diverse directions in a local region. In our patch selection, we apply the same strategy, but the window size is set to the patch size, which is an input to our program. Generally, the edge length of the patch is set to be 1/3 to 1/2 of the input blurry image. The patch size is chosen not too small for stability consideration. Nevertheless, very small patch sizes may work well in some cases and it can considerably reduce the execution time.

Fig. 1 shows some examples of synthetic data. We compare the estimated kernels computed from the whole image and some patch selection methods, including the (1) proposed method, (2) point with maximum gradient magnitude (3) central point. Some times (1) and (2) generate similar patches because corners usually have high gradient magnitude. The advantage of using (3) is that the salient part of an image usually lies in the center, and this scheme doesn't need any additional computation. Using the whole image produces the best result most of the time, but the execution time may be too long for a large image. With this effective patch selection scheme, we can find an appropriate window of a pre-selected size for the blur kernel estimation.

2.1 Multi-scale Scheme

To handle large blurs, a multi-scale optimization strategy is imperative. Generally, the initial kernel size is defined by the user. Thus this value would be easy to set if we start from a small scale. Besides, large blurs are more probable to have complex kernels than small ones. A multi-scale scheme improves the robustness and accuracy of kernel estimation.

2.2 Optimization for K

Most kernel estimation methods do not use the whole image information because there are too many noises or redundant regions that are of no use for kernel estimation. Therefore, we usually extract the edges out first.

Given a blur image patch, B_p , we solve for the kernel K by constructing the useful edge map from B_p [12]. As mentioned before, we first compute the r -map defined by

$$r(x) = \frac{\|\sum_{y \in N_h(x)} \nabla B_p(y)\|}{\sum_{y \in N_h(x)} \|\nabla B_p(y)\| + 0.5}, \quad (4)$$

where B_p denotes the blurred image patch and $N_h(x)$ is a $h \times h$ window centered at pixel x , and the constant 0.5 is to prevent producing a large r in flat regions. Then we apply thresholding on the r map to rule out pixels with small r values by

$$M = H(r - \tau_r), \quad (5)$$

where H is the Heaviside step function. The value in map M is 1 if r is higher than the threshold τ_r , and 0 otherwise. The final edge map is determined as

$$\nabla I_p^s = \tilde{\nabla} I_p \cdot H\left(M \|\tilde{\nabla} I_p\|^2 - \tau_s\right), \quad (6)$$

where \tilde{I}_p denotes the shock filtered image patch and τ_s is a threshold of the gradient magnitude. Once the edge map is obtained, we can solve the blur kernel K by minimizing the objective function

$$E(K) = \|\nabla I_p^s \otimes K - \nabla B_p\|^2 + \gamma \|K\|^2, \quad (7)$$

where γ is a weight for the regularization term. Taking the derivatives of $E(k)$ with respect to k and performing FFT on all variables, we obtain the close-form solution for K as follows:

$$K = F^{-1} \left(\frac{\overline{F(\partial_x I_p^s)} F(\partial_x B_p) + \overline{F(\partial_y I_p^s)} F(\partial_y B_p)}{F(\partial_x I_p^s)^2 + F(\partial_y I_p^s)^2 + \gamma} \right), \quad (8)$$

where F means FFT, F^{-1} means inverse FFT, and \bar{F} is the conjugate complex of F .

In our experiments, we set γ to be large, from 20 to 100, because we want the kernel to be smooth enough. If γ is too small, the kernel may break into pieces. At the end of each scale, we preserve the maximum component and leave the others out as a denoising step.

2.3 Latent Image Deconvolution

In the quadratic regularization term, the objective function of the latent image patch I_p is defined as

$$E(I_p) = \|I_p \otimes k - B_p\|^2 + \lambda \|\nabla I\|^2, \quad (9)$$

and the latent image can be solved by

$$I_p = F^{-1} \left(\frac{F(k)F(B_p)}{F(k)F(k) + \lambda(F(\partial_x)F(\partial_x) + F(\partial_y)F(\partial_y))} \right). \tag{10}$$

It is important to note that the deconvolution here is not our final deconvolution algorithm to restore the blur image. During the kernel estimation procedure, we do not need to create a well-deblurred result with complex algorithms. As long as the important edges are deconvolved well, an accurate kernel is ensured. However, this kind of simple deconvolution method produces noises in the result. We apply bilateral filtering to suppress the noises, and this step is necessary. If no denoising is performed, the noises would be amplified during the iterative procedure, thus affecting the estimated kernel.

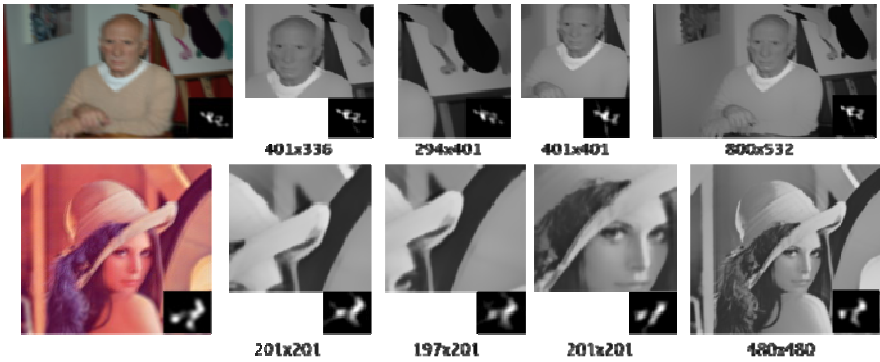


Fig. 1. Different patch selection methods and the corresponding kernel estimation. From left to right: blurred images with the ground-truth kernels, proposed method, max-gradient method, center patch, and whole image. The sizes of the patches are shown below each image.

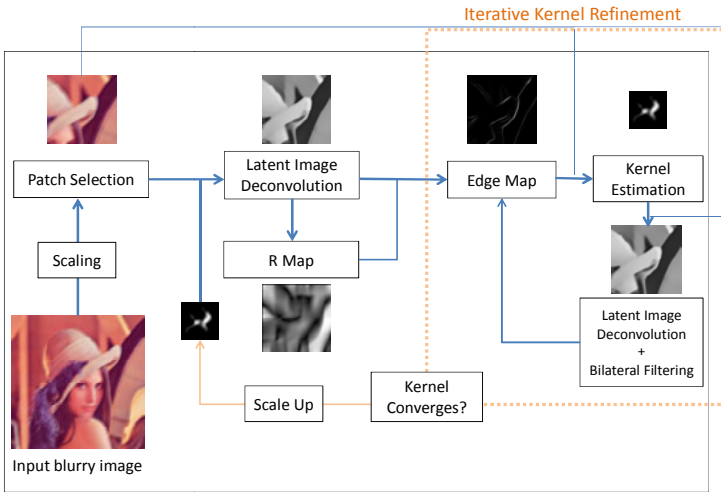


Fig. 2. Proposed framework for kernel estimation

3 Non-blind Image Deconvolution

In this paper, we propose a novel algorithm for non-blind image deconvolution, called GARL, which stands for Gradient Attenuation Richardson-Lucy. It effectively alleviates ringing artifacts by exploiting the gradient attenuation function [4] as the pixel weights to slow down the updating of pixels in the smooth regions and the regions around strong edges, thus suppressing the ringing propagation. The GARL algorithm is based on the classical Richardson-Lucy algorithm.

The Richardson-Lucy algorithm iteratively updates the image according to

$$I^{t+1} = I^t \cdot \left(F * \frac{B}{I^t \otimes F} \right), \quad (11)$$

where $*$ is the correlation operator, and t indicates the iteration number. In [7], the residual RL is proposed, which performs the RL algorithm on residual image to reduce the absolute amplitudes of the signals, hence reducing the ringing artifact. The iterative updating formula on the residual image becomes

$$\Delta I^{t+1} = (\Delta I^t + 1) \cdot \left(F * \frac{\Delta B + 1}{(\Delta I^t + 1) \otimes F} \right) - 1, \quad (12)$$

where ΔI denotes the residual image and ΔB is the residual blurred image: $\Delta B = \Delta I \otimes F + N$.

3.1 GARL

As described in [7], [9], the frequency of ringing artifact is lower than that of image details and ringing artifacts are negligible by human perception in highly textured regions. Therefore, to suppress ringing artifact, we force the smoothness constraint on the middle range of frequencies and the iterative update equation becomes

$$\Delta I^{t+1} = \frac{1}{1 + \mu W} \left\{ (\Delta I^t + 1) \cdot \left(F * \frac{\Delta B + 1}{(\Delta I^t + 1) \otimes F} \right) - 1 \right\}, \quad (13)$$

Since we want to suppress the contrast of ringing in the smooth regions while avoiding the suppression of sharp edges, the weight matrix should be large in smooth regions and small in edge and textured regions. We modify the gradient attenuation function [5] to determine the weight W for each pixel in (13). In the hierarchical restoration scheme, our modified gradient attenuation function, defined by W , is re-computed for each pixel at each scale s by propagating the scaling factor φ_s as follows:

$$\varphi_s(x, y) = \left(\frac{\alpha_s}{\|\nabla I_s(x, y)\|} \right)^\beta \cdot \left(\frac{\|\nabla I_s(x, y)\|}{\alpha_s} \right)^{\gamma M_s(x, y)}, \quad (14)$$

$$W_0 = \varphi_0, W_s = (W_{s-1})_\uparrow \cdot \varphi_s, \quad (15)$$

$$M_s = \langle \|\nabla^2 I_s\| \rangle \otimes F \cdot (1 - \langle \|\nabla I_s\| \rangle), \quad (16)$$

where $(\cdot)_\uparrow$ defines the up-sampling operator with linear interpolation, x and y denote the position in an image, α_s determines which gradient magnitude defines the smooth

regions, β and γ control the attenuating scale globally and locally, respectively, and they are set between 0.5 and 0.6 ($\beta > \gamma$) in our experiments. M_s indicates the influence range of strong edges according to the estimated blur kernel, and $\langle \cdot \rangle$ is the normalization operator. The values of M_s are between 0 and 1 so γ only effects the scaling factor at the positions where $M_s(x,y) \neq 0$. Because details are recovered more at finer scales, the gradient magnitudes of smooth area are larger than those at the same positions of the coarser scales.

3.2 Detail Recovery

The GARL can suppress most ringing artifacts with superior results compared to the other results, but it also suppresses some details around the strong edges. Therefore we propose a detail recovery process to further recover the lost details.

If we obtain two restored images from the GARL and the standard RL, the difference between them would contain the details and ringings, thus it can be expressed as follows:

$$I_{Diff} = I_{RL} - I_{GARL} = I_D + I_R \quad (17)$$

where I_{GARL} and I_{RL} denote the restored images by GARL and the standard RL, respectively, their difference is denoted by I_{Diff} , and I_D and I_R indicate the detail and ringing layer, respectively.

For each iteration, the I_R^t is obtained by applying a bilateral filter on I_{Diff}^t and then we determine a scaling factor, λ^t , to obtain a more accurate ringing layer by minimizing the following equation:

$$\lambda^t = \operatorname{argmin}_\lambda \|I_{Diff}^t - \lambda I_R^t\|^2 \quad (18)$$

$$\hat{I}_R^t = \lambda^t I_R^t \quad (19)$$

The detail layer and the difference layer for the next iteration are updated as:

$$I_D^t = \left(1 + \frac{|I_R^t|}{\max(|I_R^t|)}\right)^{-1} \cdot (1 - M) \cdot (I_{Diff}^t - \hat{I}_R^t) \quad (20)$$

$$I_{Diff}^{t+1} = I_{Diff}^t - \hat{I}_R^t - I_D^t \quad (21)$$

where M was defined in eq. (5) in the kernel estimation method.

The M map is a zero-one mask, with the pixels of value 1 representing the regions around strong edges and value 0 for smooth regions or regions containing many textures. It is inevitably to carefully select appropriate parameters to produce good restoration results. We set τ_s to be inside the range from 0.1 to 0.3, depending on the characteristics of the images.

The final restored image, I_F , is determined by:

$$I_F = I_{GARL} + \sum_t I_D^t \quad (22)$$

Fig.3 is an example of GARL method. We also compare the results with other methods [7], [8]. The GARL result apparently contains less ringing artifact. Fig.4 shows the procedure of the detail recovery procedure.

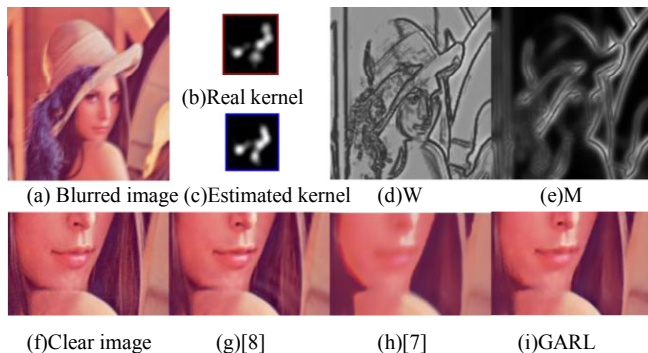


Fig. 3. An example of GEARL algorithm compared with other methods

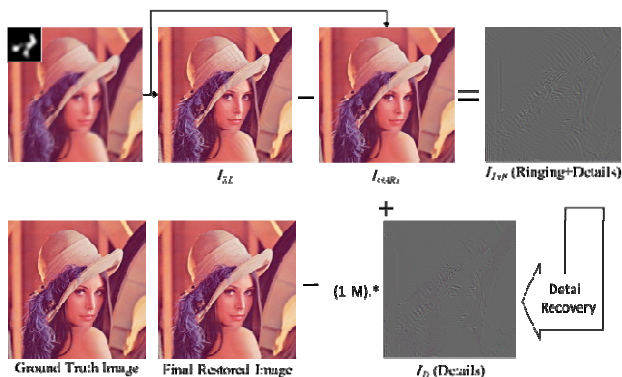




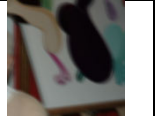




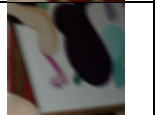





Fig. 4. Flowchart of the proposed detail recovery procedure

4 Experimental Results

In this section, we show some quantitative performance of our patch selection method and some experimental results to demonstrate the ringing suppression by using the proposed algorithm. Also, we provide some deblurred results on real images. Real images contain many unexpected factors, such as spatial varying blur or noises, which may lead to failed cases of kernel estimation or deconvolution. The proposed method, however, can restore most real blurred images well with the ringings suppressed and the details recovered. The computing platform for our experiments is a notebook running MS Vista 32-bit version with Intel Core2 CPU P7450 2.13GHz, and 4GB RAM. The program was implemented in MATLAB.

We measure the performance of our patch selection method by computing the minimum SAD between the estimated kernels and the ground truth. From Table 1, our Harris corner patch selection method produces the kernel which is most similar to the ground truth. Although the patches of scale 3 in method (1) and (2) look similar, the

Table 1. Performance of the patch selection methods. (1)proposed Harris corner patch method, (2)patch with maximum gradient magnitude, (3)central patch. The number below each estimated kernel is the minimum SAD value with the ground truth kernel.

Method	Scale 1	Scale 2	Scale 3	kernel	GT
(1)				 0.4218	
(2)				 0.6851	
(3)				 0.4254	

estimated kernels differs because (2)’s kernel is not estimated well in scale 2, affecting the result. Therefore, finding an appropriate patch is very important in every scale.

Some detailed deblurred results are shown in Fig. 5 and Fig. 6. These experimental comparisons demonstrate the improved image restoration results by using the proposed algorithm, especially in the ringing suppression and detail recovery. The revised detail recovery results contain more texture than GARL and less ringing than the previous Richardson-Lucy based deconvolution methods. We compare our results with some other representative previous works [9][14][15] in Fig. 7. More well-deblurred results by using the proposed algorithm on real images are shown in Fig. 8 and Fig. 9.

As for the execution time, our program also performs well. We have tried two methods to solve the objective functions given in eq. (7) and (9), as in Table.2. Using conjugate gradient method for solving eq. (7) and RL method for eq. (9) generates more controllable and stable results. However, the FFT division method considerably decreases the execution time. This experiment is tested on a 900x900 image (statue).

Table 2. Execution time of the proposed image deblurring algorithm

Objective function (7)	Objective function (9)	Execution time((7)+(9))
Conjugate gradient	Richardson-Lucy	330s
FFT division	FFT division	46s

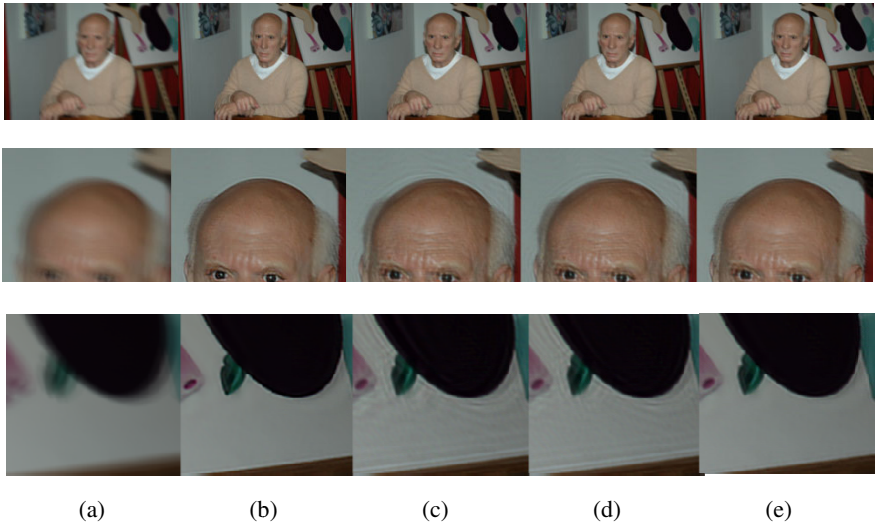


Fig. 5. Detail recovery example 1. (a) blurred image, restored images by using (b) GEARL, (c) standard RL, (d) detail recovery, and (e) revised detail recovery.

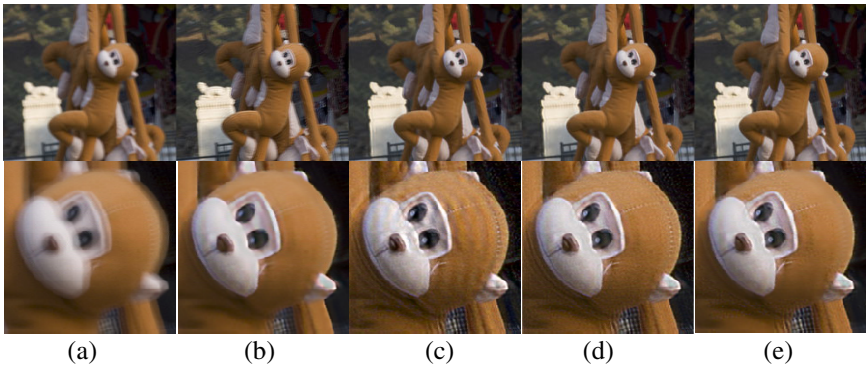


Fig. 6. Detail recovery example 2. (a) Blurred image. (b) GEARL. (c) standard RL. (d) Detail recovery. (e) Revised detail recovery.

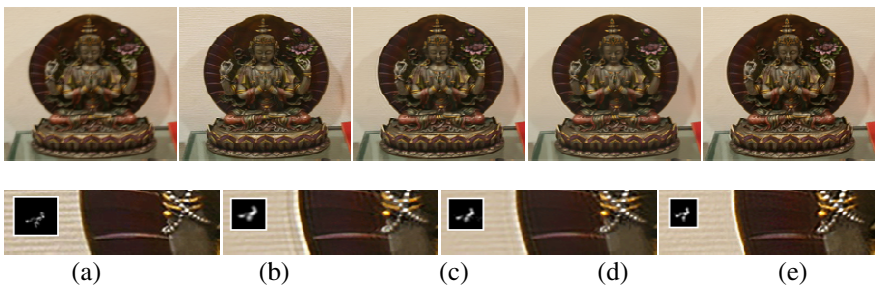


Fig. 7. (Top) (a) Blurred image, deblurred images by using (b) Shan et al. [9], (c) Joshi et al. [14], (d) Levin et al. [15], and (e) the proposed deblurring algorithm. (Down) Enlarged parts of (b)~(e).



Fig. 8. (a) A real blurred image and (b) the deblurred image by using the proposed deblurring algorithm

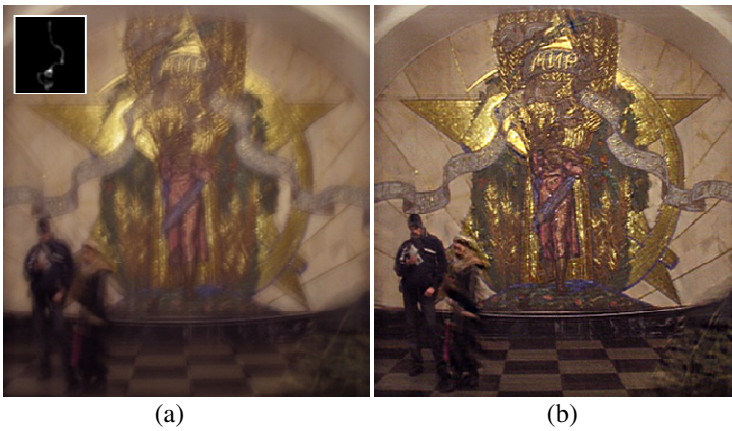


Fig. 9. (a) A real blurred image and (b) the deblurred image by using the proposed deblurring algorithm

5 Conclusion

In this paper, we proposed a framework for image deblurring from a single blurred image. The proposed blur kernel estimation can effectively suppress the ringing artifact and recover the image details compared to the previous methods. However, the proposed image deblurring algorithm has the following limitations. First, using FFT to speed up the computation in the kernel estimation procedure may produce some problem near the image boundary. In addition, like many other deblurring methods, the proposed algorithm also requires the user to carefully tune the parameters to obtain the optimized deblurred image. In the future, developing a fully automatic image deblurring system is the direction of our research.

References

1. Richardson, W.H.: Bayesian-based iterative method of image restoration. *Journal of the Optical Society of America* 62, 55–59 (1972)
2. Lucy, L.B.: An iterative technique for the rectification of observed distributions. *Astronomical Journal* 79, 745–765 (1974)
3. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: *ICCV*, pp. 839–847 (1998)
4. Fergus, R., Singh, B., Hertzmann, A., Roweis, S.T., Freeman, W.T.: Removing camera shake from a single photograph. *ACM Trans. Graphics* 25, 787–794 (2006)
5. Fattal, R., Lischinski, D., Werman, M.: Gradient domain high dynamic range compression. *ACM Trans. Graphics*, 249–256 (2002)
6. Levin, A.: Blind motion deblurring using image statistics. In: *NIPS*, pp. 841–848 (2006)
7. Levin, A., Fergus, R., Durand, F., Freeman, W.T.: Image and depth from a conventional camera with a coded aperture. *ACM Trans. Graphics* 26, 70 (2007)
8. Yuan, L., Sun, J., Quan, L., Shum, H.-Y.: Progressive inter-scale and intra-scale non-blind image deconvolution. *ACM Trans. Graphics* 27, 1–10 (2008)
9. Shan, Q., Jia, J., Agarwala, A.: High-quality motion deblurring from a single image. *ACM Trans. Graphics* 27, 73–83 (2008)
10. Yuan, L., Sun, J., Quan, L., Shum, H.-Y.: Image deblurring with blurred/noisy image pairs. *ACM Trans. Graphics* 26 (2007)
11. Cho, S., Lee, S.: Fast motion delurring. *ACM Trans. Graphics (SIGGRAPH ASIA)* (2009)
12. Xu, L., Jia, J.: Two-Phase Kernel Estimation For Robust Motion Deblurring. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010. LNCS*, vol. 6311, pp. 157–170. Springer, Heidelberg (2010)
13. Harris, C., Stephens, M.: A Combined corner and edge detector. In: *Alvey Vision Conference* (1988)
14. Joshi, N., Szeliski, R., Kriegman, D.: PSF estimation using sharp edge prediction. In: *CVPR* (2008)
15. Levin, A., Sand, P., Cho, T.S., Durand, F., Freeman, W.T.: Motion-invariant photography. In: *SIGGRAPH* (2008)

Quality Estimation for H.264/SVC Inter-layer Residual Prediction in Spatial Scalability

Ren-Jie Wang¹, Yan-Ting Jiang¹, Jiunn-Tsair Fang², and Pao-Chi Chang¹

¹Dept. of Communication Engineering, National Central Univ., Jhongli, Taiwan

²Dept. of Electronic Engineering, Ming Chuan Univ., Taoyuan, Taiwan

{rjwang, ytjiang}@vaplab.ce.ncu.edu.tw, fang@mail.mcu.edu.tw,
pcchang@ce.ncu.edu.tw

Abstract. Scalable Video Coding (SVC) provides an efficient compression for the video bitstream equipped with various scalable configurations. H.264 scalable extension (H.264/SVC) is the most recent scalable coding standard. It involves the state-of-the-art inter-layer prediction to provide higher coding efficiency than previous standards. Moreover, the requirements for the video quality on distinct situations like link conditions or video contents are usually different. Therefore, it is very desirable to be able to construct a model so that the target quality can be estimated in advance. This work proposes a Quantization-Distortion (Q-D) model for H.264/SVC spatial scalability, and then we can estimate video quality before the actual encoding is performed. In particular, we further decompose the residual from the inter-layer residual prediction into the previous distortion and Prior-Residual so that the residual can be estimated. In simulations, based on the proposed model, we estimate the actual Q-D curves, and its average accuracy is 88.79%.

Keywords: H.264, Scalable Video Coding, Spatial Scalability, Quality Estimation, Quantization-Distortion Model.

1 Introduction

The fundamental principle of Scalable Video Coding (SVC) is to generate a single compressed bit stream that can adapt to the varying bit rates, display resolutions, and computational resource constraints of various receivers rapidly and easily. There are three kinds of scalability, including temporal, spatial, and quality (SNR) scalability. The spatial scalability that provides various resolutions is suitable for display devices with different sizes nowadays available. In order to remove redundancy between layers, the enhancement layer can be coded using the inter-layer prediction which includes the motion, texture and residual information from the base layer. In H.264/SVC, there exist three kinds of inter-layer prediction tools. There are Inter-Layer Motion Prediction (ILMP), Inter-Layer Intra Prediction (ILIP), and Inter-Layer Residual Prediction (ILRP) [1][2]. ILMP up-samples motion vectors as a motion predictor. ILIP up-samples the reconstructed blocks for the prediction of luminance. Moreover, ILRP up-samples the residual for the residual compensation.

The requirement for the video quality on distinct situations like link conditions or video content is usually different. Therefore, it is very desirable to be able to construct a Quantization Distortion (Q-D) model so that the target quality can be achieved by selecting a proper encoder Quantization Parameter (QP). Most of the proposed Q-D models were for a single layer video coding [3-7]. In particular, their models were based on the assumption of residual distributions [3-5]. That is, the distortion can be modeled as a function of QP and the variance of the residual distribution. Recently, two Q-D models for SVC spatial scalability and temporal scalability were proposed to perform the optimal rate allocation [8][9]. For the real time application, their parameters of the model are estimated during the encoding procedure.

In this work, we propose a Q-D model for H.264/SVC spatial scalability to estimate video quality. However, the model parameter and the quality score have to be obtained before the entire coding procedure starts. We introduce a residual decomposition technique for ILRP, in which the residual can be decomposed to the coding error and the displacement difference (Prior-Residual). Then the distortion can be modeled as a function of quantization step and Prior-Residual that can be estimated before encoding.

In the remaining of this paper, the analysis of the distortion in the transform domain and related works on Q-D model are discussed in Section 2. The proposed Q-D model and quality estimation for ILRP are described in Section 3. The results for validating the accuracy of proposed model and specifying the model parameters are depicted in Section 4. Finally we summarize our proposed method and results in conclusion.

2 Distortion Analysis and Related Works

2.1 Distortion Analysis in the Transform Domain

Most literatures on Q-D modeling analyze the distortion, specifically the Mean Square Error (MSE) between the original and the reconstructed frames, in the transform domain [3][4]. Two major reasons are that transform coefficients have more similar characteristics than pixels in the spatial domain among various video contents, and the quantization in hybrid video coding, which is the basic structure for most current video coding standards, is performed in the transform domain.

From Fig. 1, we observe that the difference between original frame f_k and reconstructed frame $f'_k(q)$ equals to the difference between residual $r_k(q)$ and quantized residual $r_k^{quan}(q)$ as shown in (1).

$$\begin{aligned} f_k - f'_k(q) &= [f_k - MC(f'_{k-1}(q))] - [f'_k(q) - MC(f'_{k-1}(q))] \\ &= r_k(q) - r_k^{quan}(q) \end{aligned} \quad (1)$$

Because DCT transform is linear, the equality holds in DCT domain.

$$F_k - F'_k(q) = R_k(q) - R_k^{quan}(q) \quad (2)$$

By Parseval's theorem,

$$E[(f_k - f'_k(q))^2] = E[(F_k - F'_k(q))^2] = E[(R_k(q) - R_k^{quan}(q))^2] \quad (3)$$

we can derive that the MSE between the original and the reconstructed frames equals to the MSE between the original and the quantized residuals in the transform domain. Hence, with the assumption for residual distribution, the distortion is possible to be modeled as a function of QP and parameters of the distribution.

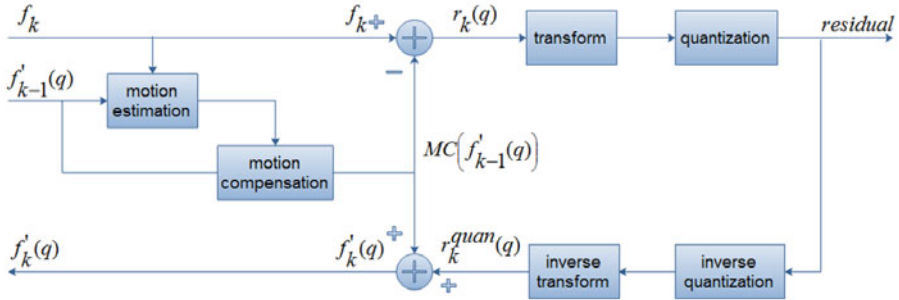


Fig. 1. DPCM based encoder structure

2.2 Laplacian and Cauchy Distributions for DCT Coefficients

The quantized residual can be regarded as a Laplacian-distributed random variable [3]. Then a closed-form expression of distortion is derived. Recently, Kamaci *et al.* [4] proposed Cauchy density function as the residual distribution. Its probability density function (pdf) is shown as

$$p(x) = \frac{1}{\pi} \frac{\mu}{\mu^2 + x^2} \quad (4)$$

where μ is the half-width at half-maximum of the pdf. It basically reflects the variance of the distribution, and can be denoted as a function of σ_x^2 , *i.e.*, $\mu = h(\sigma_x^2)$.

The closed-form expression of the distortion is derived and approximated to a power function of q in [4] as

$$\begin{aligned}
 D(q) &= \sum_{i=-\infty}^{\infty} \int_{(i-\frac{1}{2})q}^{(i+\frac{1}{2})q} (x-iq)^2 p(x) dx \\
 &= \sum_{i=-\infty}^{\infty} \int_{(i-\frac{1}{2})q}^{(i+\frac{1}{2})q} (x-iq)^2 \frac{1}{\pi} \frac{h(\sigma_x^2)}{(h(\sigma_x^2))^2 + x^2} dx \approx aq^b = f(\sigma_x^2, q)
 \end{aligned} \quad (5)$$

where $a, b > 0$, and depend on σ_x^2 .

It also demonstrates the Cauchy density is more accurate in estimating the distribution of the DCT coefficients than the traditional Laplacian density. Furthermore, it yields less estimation error for Q-D curve. Therefore, Cauchy distribution is assumed in our work.

2.3 Residual Decomposition for Single Layer

For residual decomposition, Guo *et al.*[10] proposed a quality estimation method for single layer coding. The residual can be decomposed into the displacement difference and the coding error as shown in (6).

$$\begin{aligned} R_k(q) &= F_k - F'_{k-1}(q) \\ &= (F_k - F'_{k-1}) + (F'_{k-1} - F'_{k-1}(q)) \\ &= I_k + E_{k-1}(q) \end{aligned} \quad (6)$$

The residual $R_k(q)$ is the difference between the original frame F_k and the predicted frame $F'_{k-1}(q)$ that is compensated for by the previous reconstructed frame. On the other hand, F'_{k-1} is the predicted frame that is compensated for by the previous original frame. The residual can be decomposed to the displacement difference I_k , and the coding distortion of the previous frame $E_{k-1}(q)$. Furthermore, with the assumption that both I_k and $E_{k-1}(q)$ have zero mean and are uncorrelated, the variance is also decomposable as (7) shows. That is, the variance $\sigma_{R_k}^2(q)$ is equal to the sum of $\sigma_{I_k}^2(q)$ and $\sigma_{E_{k-1}}^2(q)$.

$$\sigma_{R_k}^2(q) = \sigma_{I_k}^2 + \sigma_{E_{k-1}}^2(q) \quad (7)$$

3 Proposed Q-D Estimation Method

In this section, we describe the proposed Q-D estimation method for H.264/SVC inter-layer residual prediction in spatial scalability in detail. With the power form Q-D model and the residual decomposition as basis, we can build up the quality estimation mechanism. The Q-D model for single layer coding that prediction data only come from its own layer is described first. The model can be applied to the base layer or enhancement layers without inter-layer prediction in SVC. Moreover, for SVC inter-layer prediction, the enhancement layer quality or the residual will vary with the similarity between two layers. A Q-D mode for enhancement layers with inter-layer residual prediction is then proposed.

3.1 Q-D Model for Single Layer Coding

As mentioned in [10], with the assumption that a video sequence is a locally temporal stationary process, the corresponding variables in successive frames have the same variance. Thus

$$\begin{aligned} \sigma_R^2(q) &= \sigma_{I_k}^2 + \sigma_{E_{k-1}}^2(q) \\ &= \sigma_I^2 + \sigma_E^2(q) \\ &\stackrel{\Delta}{=} PR + D(q) \end{aligned} \tag{8}$$

where Prior-Residual, defined as $PR = \sigma_I^2$, is the variance of the displacement difference.

Then, we can put (8) into (5) to obtain (9)

$$\begin{aligned} D(q) &= f(\sigma_R^2(q), q) \\ &= f(PR + D(q), q) \end{aligned} \tag{9}$$

Because $D(q)$ is a function of PR , and the DCT coefficients are modeled by Cauchy distribution in [4], (9) can be further simplified as (10).

$$D(q) = f'(PR, q) \approx aq^b = aq^{cPR^d} \tag{10}$$

Where a , c , and d are constants. The specific relationship between b and PR can be built up by empirical tests. We will observe that PR can accurately predict the distortion curve as a good parameter to identify the sequence characteristic.

3.2 Q-D Model for Inter-layer Residual Prediction

The inter-layer residual prediction in SVC is depicted as Fig. 2. Because the high correlation of residual signals between the current and the reference layer, the difference $R_{RP_k}(q)$ between the residuals of two layers instead of residual signal itself is encoded as the enhancement information to improve the coding efficiency.

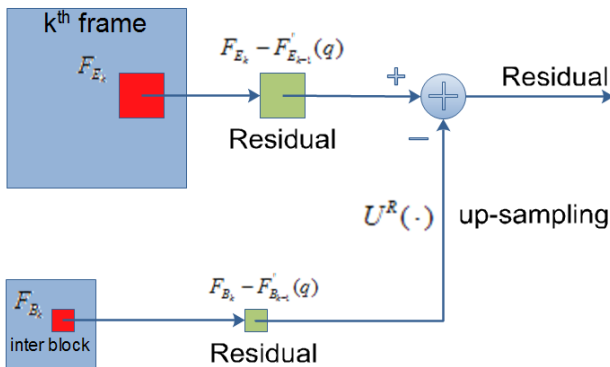


Fig. 2. Inter-layer residual prediction structure in SVC spatial scalability

We also employ the residual decomposition to this structure. By involving the predicted frame by non-distorted data in the enhancement layer $F'_{E_{k-1}}$ and that for the base layer $F'_{B_{k-1}}$, the residual can also be decomposed to the distortion from imperfect prediction and quantization error as in (11).

$$\begin{aligned}
 R_{RP_k}(q) &= F_{E_k} - F'_{E_{k-1}}(q) - U^R(F_{B_k} - F'_{B_{k-1}}(q)) \\
 &= (F_{E_k} - F'_{E_{k-1}}) + (F'_{E_{k-1}} - F'_{E_{k-1}}(q)) \\
 &\quad - U^R((F_{B_k} - F'_{B_{k-1}}) + (F'_{B_{k-1}} - F'_{B_{k-1}}(q))) \\
 &= I_{E_k} + E_{E_{k-1}}(q) - U^R(I_{B_k} + E_{B_{k-1}}(q)) \\
 &= [I_{E_k} - U^R(I_{B_k})] + [E_{E_{k-1}}(q) - U^R(E_{B_{k-1}}(q))]
 \end{aligned} \tag{11}$$

where $U^R(\cdot)$ means the upsampling procedure, which can be implemented by simple bi-linear interpolation or any more sophisticate interpolation operations.

We assume that both $I_{E_k} - U^R(I_{B_k})$ and $E_{E_{k-1}}(q) - U^R(E_{B_{k-1}}(q))$ have zero mean and are approximately uncorrelated. In addition, a video sequence is a locally temporal stationary process, *i.e.*, the corresponding variables in consecutive frames have the same variance. (11) can be derived as (12).

$$\begin{aligned}
 \sigma_{R_{RP}}^2(q) &= \text{var}([I_{E_k} - U^R(I_{B_k})] + [E_{E_{k-1}}(q) - U^R(E_{B_{k-1}}(q))]) \\
 &= \text{var}([I_{E_k} - U^R(I_{B_k})]) + \text{var}(E_{E_{k-1}}(q)) + \text{var}(U^R(E_{B_{k-1}}(q))) \\
 &\quad - 2\rho\sqrt{\text{var}(E_{E_{k-1}}(q))}\sqrt{\text{var}(U^R(E_{B_{k-1}}(q)))} \\
 &= PR_{RP} + (1 + \beta - 2\rho\sqrt{\beta}) \cdot D_{RP}(q)
 \end{aligned} \tag{12}$$

where Prior-Residual for Residual Prediction is denoted as $PR_{RP} = \text{var}(I_{E_k} - U^R(I_{B_k}))$. Because of high dependency between two layer distortions, the base layer distortion can be predicted by the enhancement one, which is $\text{var}(U^R(E_{B_{k-1}}(q))) = \beta \text{var}(E_{E_{k-1}}(q)) = \beta \cdot D_{RP}(q)$. For most video sequences, $\beta > 1$ since the higher residual variance exist in the downsampled frame. β and ρ , which is the correlation coefficient between $E_{E_{k-1}}(q)$ and $U^R(E_{B_{k-1}}(q))$, vary only slightly with different video contents and hence can be consider as constants. The distortion term $\text{var}(E_{E_{k-1}}(q) - U^R(E_{B_{k-1}}(q)))$ can be expressed as a constant times of $D_{RP}(q)$. Therefore, it is possible to use PR_{RP} to predict the real residual before encoding procedure including Rate Distortion Optimization (RDO), transform, and quantization procedure.

Then, we can put (12) into (5) to obtain (13).

$$\begin{aligned}
 D_{RP}(q) &= f(\sigma_{R_{RP}}^2(q), q) \\
 &= f(PR_{RP} + (1 - \beta - 2\rho\sqrt{\beta}) \cdot D_{RP}(q), q)
 \end{aligned}
 \tag{13}$$

Because $D_{RP}(q)$ is a function of PR_{RP} , and the DCT coefficients are modeled by Cauchy distribution, (13) can be further simplified as (14).

$$D_{RP}(q) = f'(PR_{RP}, q) \approx aq^b = aq^{cPR_{RP}^d}
 \tag{14}$$

where b can also be represented by a power form with PR_{RP} . It will be verified with c and d in the experiment section with real video data. Note that, PR_{RP} means the Prior-Residual for Residual Prediction, which is different from PR for single layer, and it provides more accurate description for Q-D behavior.

Block diagram for obtaining PR_{RP} is shown in Fig.3. Based on the obtained PR , Q-D function for a certain video sequence is established. We then can either predict the distortion according to a given QP, or select a suitable QP to obtain the video with target visual quality.

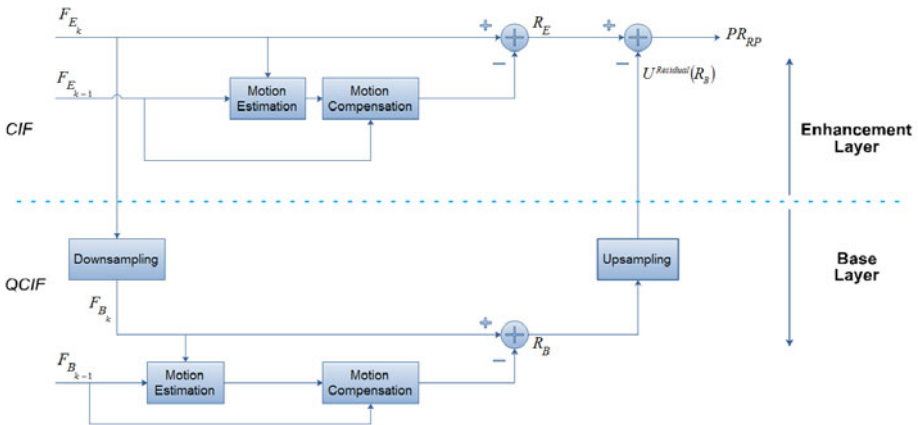


Fig. 3. Prior-Residual in inter-layer residual prediction

4 Experimental Results

In this section we construct an experiment to verify the proposed distortion model for ILRP. We will obtain the model parameters by fitting real coding results in the training phase, and then the performance with those sequences outside the training set will be demonstrated. Experiment setting is the following. Four training video sequences for two layers in CIF and QCIF formats at the frame rate of 30 frames/s, including Akiyo, Carphone, Harbour, Mobile are encoded by H.264/SVC reference software

JSVM 9.19.8. Six QPs (16, 20, 24, 28, 32, 36) are used in the encoding. The same QPs are used for both the base layer and the enhancement layer. We used 90 frames for training, and the first frame is an I-frame while the rest are P-frames. The inter-layer prediction flag was the inter-layer residual prediction (0,0,1). Five test video sequences including Bus, Foreman, Hall, Mother_daughter, and Soccer are encoded, and the rest experiments setting are the same with training process.

As Fig. 4 shows, black dots are the results after the SVC coding for four training sequences. The dotted lines are the approximated curves based on the power form Q-D relationship. We can obtain the specific value b that minimizes the estimate distortion for each sequence in the Table 1. Note that we pre-set a as a constant to simplified the model. As shown in Table 1, the numerical value of b can reflect the behavior that higher distortion or complicated content has a larger b at same QP among different video contents.

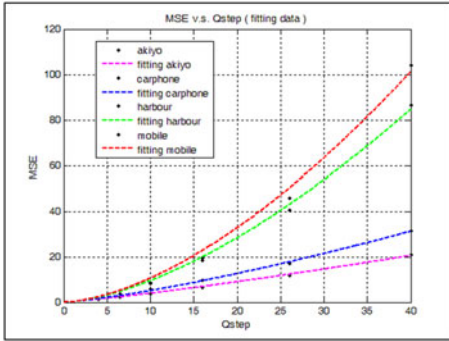


Fig. 4. The training Q-D curve in inter-layer residual prediction

Table 1. The Q-D model parameter in inter-layer residual prediction

$$MSE = a * Qstep ^ b$$

$a * x^b$	a	b	R ²
akiyo	0.254	1.191	0.99
carphone	0.254	1.304	0.99
harbour	0.254	1.576	0.99
mobile	0.254	1.624	0.99

From the experimental results, the Q-D model of inter-layer residual prediction can be precisely specified as the following

$$D(q) = MSE = 0.254 * Qstep^b \quad (15)$$

As we derived in the Section 3, the constant b is only related to the residual variance or PR . Hence, we observe the relationship between b and PR by Fig. 5. From the training data, represented by blue squares in the figure, we can observe that b can be modeled as a power function of PR , and c and d can be determined to be 1.03 and 0.08, respectively. The determination coefficient R^2 in the fitting process is up to 0.97, which implies an excellent fitting.

From the empirical data, the relationship between b and PR_{RP} in inter-layer residual prediction is shown as

$$b = 1.037 * PR_{RP}^{0.080} \quad (16)$$

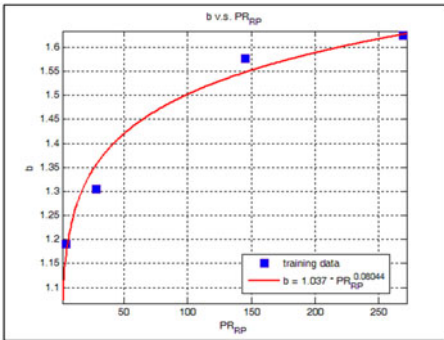


Fig. 5. The fitting curve about b and PR_{RP} in inter-layer residual prediction

Table 2. The relationship between b and PR_{RP} in inter-layer residual prediction

training	b	PR_{RP}
akiyo	1.191	4.18
carphone	1.304	28.16
harbour	1.576	145.50
mobile	1.624	269.51

fitting function	R^2
$b = 1.037 * PR_{RP}^{0.080}$	0.97

Fig.6 and Fig. 7 show the real and the estimated Q-D curves, respectively. It clearly shows that the estimate curves can fit the results obtaining from time-consuming SVC coding. Accuracy of the proposed Q-D model, defined as in (17), for various sequences and Qsteps are listed in Table 3. The average accuracies for all test sequences are more than 81.19%, and up to 93.54% in the sequence Hall. Translated to PSNR, the estimation error is no more than 0.74 dB.

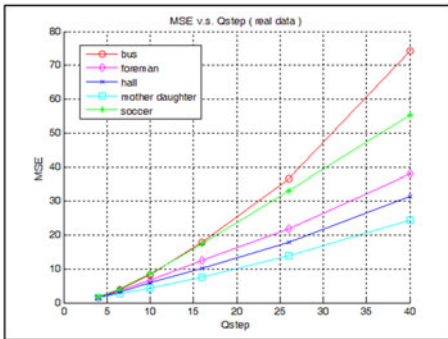


Fig. 6. The encoded Q-D curve in inter-layer residual prediction

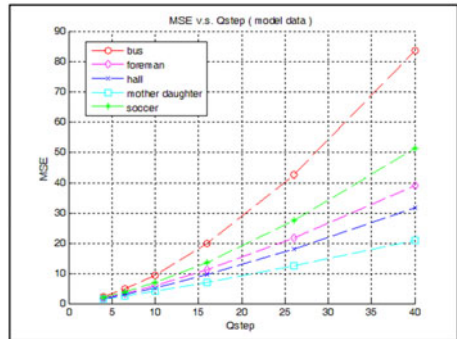


Fig. 7. The modeled Q-D curve in inter-layer residual prediction

Table 3. The accuracy of the Q-D model in inter-layer residual prediction

EL with ILRP cif	Qstep						Average Accuracy
	4	6.5	10	16	26	40	
bus	62.06	76.81	87.16	89.38	84.25	87.49	81.19
foreman	94.85	90.86	87.23	89.55	98.83	97.54	93.14
hall	91.98	91.67	86.09	93.23	98.99	99.31	93.54
mother daughter	91.73	91.98	91.49	90.87	88.80	85.58	90.07
soccer	88.36	92.38	81.28	78.56	83.30	92.22	86.02

$$\text{Accuracy} = \left(1 - \frac{|\text{Actual MSE} - \text{Estimated MSE}|}{\text{Actual MSE}} \right) \times 100\% \quad (17)$$

5 Conclusion

We have proposed a Q-D model for inter-layer residual prediction in SVC. The distortion is modeled as a function of quantization step and Prior-Residual that can be efficiently estimated before encoding. Experimental results show that the proposed model can estimate the actual Q-D curves for inter-layer prediction, and the average accuracy of the model is 88.79% in MSE or the estimated error less than 0.74 dB in PSNR, which is suitable for practical use. In the future, we will extend the residual decomposition and Q-D modeling to all inter-layer prediction tools.

References

1. Schwarz, H., Marpe, D., Wiegand, T.: Overview of the Scalable Video Coding Extension of the H.264/AVC Standard. *IEEE Trans. Circuits Syst. Video Technol.* 17(9), 1103–1120 (2007)
2. Segall, A., Sullivan, G.J.: Spatial Scalability Within the H.264/AVC Scalable Video Coding Extension. *IEEE Transactions on Circuits and Systems for Video Technology* 17(9), 1121–1135 (2007)
3. Turaga, D.S., Chen, Y., Caviedes, J.: No reference PSNR estimation for compressed pictures. *Signal Process. Image Commun.* 19, 173–184 (2004)
4. Kamaci, N., Altinbasak, Y., Mersereau, R.M.: Frame bit allocation for the H.264/AVC video coder via Cauchy density-based rate and distortion models. *IEEE Trans. Circuits Syst. Video Technol.* 15(8), 994–1006 (2005)
5. Berger, T.: *Rate-Distortion Theory: A Mathematical Basis for Data Compression*. Prentice-Hall, Englewood Cliffs (1971)
6. Takagi, K., Takishima, Y., Nakajima, Y.: A study on rate distortion optimization scheme for JVT coder. In: *Proc. SPIE*, vol. 5150, pp. 914–923 (2003)
7. Wang, H., Kwong, S.: A rate-distortion optimization algorithm for rate control in H.264. In: *Proc. IEEE ICASSP 2007*, pp. 1149–1152 (April 2007)
8. Liu, J., Cho, Y., Guo, Z., Kuo, C.C.: Bit Allocation for Spatial Scalability Coding of H.264/SVC With Dependent Rate-Distortion Analysis. *IEEE Trans. Circuits Syst. Video Technol.* 20(7), 967–981 (2010)
9. Hu, S.H., Wang, H., Kwong, S., Zhao, T., Kuo, C.C.: Rate Control Optimization for Temporal-Layer Scalable Video Coding. *IEEE Trans. Circuits Syst. Video Technol.* 21(8), 1152–1162 (2011)
10. Guo, L., Au, O.C., Ma, M., Liang, Z., Wong, P.H.W.: A Novel Analytic Quantization-Distortion Model for Hybrid Video Coding. *IEEE Trans. Circuits Syst. Video Technol.* 19(5), 627–641 (2009)

Extracting Interval Distribution of Human Interactions

Ryohei Kimura¹, Noriko Takemura¹, Yoshio Iwai², and Kosuke Sato¹

¹ Graduate School of Engineering Science, Osaka University

{kimura, takemura}@sens.sys.es.osaka-u.ac.jp

Sato@sys.es.osaka-u.ac.jp

² Graduate School of Engineering, Tottori University

Iwai@ike.tottori-u.ac.jp

Abstract. Recently, activity support systems that enable dialogue with humans have been intensively studied owing to the development of various sensors and recognition technologies. In order to enable a smooth dialogue between a system and a human user, we need to clarify the rules of dialogue, including how utterances and motions are interpreted among human users. In conventional study on dialogue analysis, duration between the time when someone finishes an utterance and the time when another human starts the next utterance were analyzed. In a real dialogue between humans, however, there are sufficient intervals between an utterance and a visually observable motion such as bowing and establishing eye-contact; the facilitation of communication and cooperation seem to depend on these intervals. In our study, we analyze interactions that involve utterances and motions at a reception scenario by resolving motions into motion primitives (a basic unit of motion). We also analyze the timing of utterances and motions in order to structure dialogue behaviors. Our result suggest that a structural representation of interaction can be useful for improving the ability of activity support systems to interact and support human dialogue.

Keywords: Structural Representation of Interaction, Action Primitives, Interval Analysis.

1 Introduction

Recently, the development of various sensors and image and speech recognition technologies have fostered more intensive studies of systems that support dialogue with human [1,2,3]. Although interaction among humans or between a human and a system are diversifying, as in the systems mentioned above, it is difficult for humans to communicate with these systems as smoothly as they would in a daily face-to-face human conversation. For example, when humans interpret the thoughts of a dialogue partner, they holistically use not only aural information such as the context of utterance but also visual information such as body language including a nod, a gesture, or lip movements. In addition, during human conversation, it is important for one partner to start talking before or

soon after the other partner finishes his or her utterance in order to communicate that the intent of utterance was understood. In conventional systems, the information for predetermined demands from users are communicated to other users and they cannot adjust them [4,5]. Redundant intervals between utterances occur because of delays in communication and low computational speed, and it becomes difficult to communicate the intent of the utterance correctly. In such situations, incorrect recognition of the intent of the utterance or a delay in the response from a dialogue partner can cause anxiety, because the partners cannot understand the intention and the utterances from each partner sometimes collide owing to redundant utterances. Therefore, in conventional study on dialogue analysis, various human interactions were analyzed to clarify the relationship between a speaker's intent and the structure of the interaction. However, tagging of speech intentions or evaluating temporal structures is done by one of the interacting participants or a third party and depends heavily on their subjective judgments. Therefore, in our study, we extract structural representations automatically from speech and visual information such as body movements during human interaction.

The rest of the paper is organized as follows. Section 2 presents a comparative analysis between our study and related studies, and section 3 describes the recognition model of conversation behavior. Section 4 describes the database of interaction behavior compiled in a register that is used at the reception in the experiment, and section 5 describes our experimental results showing extracted response time distribution. Section 6 summarizes the paper and discusses future works.

2 Related Works

Kawashima et al. analyzed the temporal structure of head movements and utterances in *Rakugo* and extracted the timing when the performers switch roles by using visible gestures. Although they analyzed the temporal structure of multi modal interactions, such transitions are not represented structurally [9].

Sumi et al. proposed a structural representation of human interactions by using multi modal data obtained by using various sensors; this approach is similar to the used in the present study [10]. In that study, a three-party conversation in a poster presentation was represented by three annotations: utterance information, eye directions, and pointing directions. The study assumed that the transitions among speakers in the three-party conversation were affected by only the previous state and could be represented by a tree structure using an n-gram model. However, a semantic-level representation such as an annotation depends on the interpretation of individuals and requires manual extraction from sensor data. It is, therefore, difficult to extract a semantic-level representation automatically. Moreover, an n-gram model cannot express the duration of the dialogue, which plays an important role during an interaction, because the symbols used in an n-gram model express only the state of the dialogue.

In this paper, we model dialogue behavior by using action primitives, a minimal unit of action, which can be automatically extracted at the signal level from

sensor data, such as three-dimensional (3-D) joint positions or voice signals, and do not require manual annotations. The temporal relationship between dialogue partners is represented by the transition probability of dialogues learned from the intervals of the action primitives extracted from sensor data that were used as training data. Since a dialogue is modeled at a signal level and not at a semantic level, we can make a representation of the implicit information of dialogue partners, which cannot be achieved at a semantic-level.

3 Dialogue Recognition Model

In this section, the recognition model of dialogue behavior between humans are described.

3.1 Structuring the Dialogue Behavior

Fig. 1 shows how the model of dialogue behavior was structured for using in the present study. In our study, a dialogue behavior is represented by a motion primitive, m , which is a minimum unit of motion extracted automatically from multi modal data. The detailed method used to extract motion primitives is described in Sec.4. Extracted motion primitives are denoted by m_1, m_2, \dots, m_N in the order of observation, and the start time and finish time of m_i are denoted by t_s^i and t_e^i respectively. We define a basic dialogue behavior I_S as an interactional subsequence, i.e., a prior motion primitive of an asker $m_a \rightarrow$ recognizing m_a by a responder \rightarrow a posterior motion primitive of the responder $m_r \rightarrow$ recognizing m_r by the asker. Sequential dialogue behavior is expressed as the transition of the basic dialogue behavior. The posterior motion primitive m_r^i in the i -th basic dialogue behavior I_S^i is equivalent to the prior motion primitive m_a^{i+1} in the $i + 1$ -th basic dialogue behavior I_S^{i+1} . Human dialogue behaviors are defined as a chain model of following three probabilities: p_0 , p_1 , and p_2 . For illustrative purposes, it is assumed that human A is an asker and human B is a responder in the basic dialogue behavior I_S^i .

- p_0 : the probability of an initial motion primitive
 p_0 is the probability of observing the motion primitive m_a^1 by human A when no motion primitives were observed right before that time. p_0 is defined as follows:

$$p_0^A(m_a^1) \quad (1)$$

- p_1 : the probability of transiting motion primitives ($m_a \rightarrow m_r$)
 p_1 is the probability of starting the motion primitive m_r^i by human B at time t_s^{i+1} when the motion primitive m_a^i finish at time t_e^i in the basic dialogue behavior I_S^i . p_1 is defined as follows:

$$p_i^B(m_r^i, t_s^{i+1} | m_a^i, t_e^i) \quad (2)$$

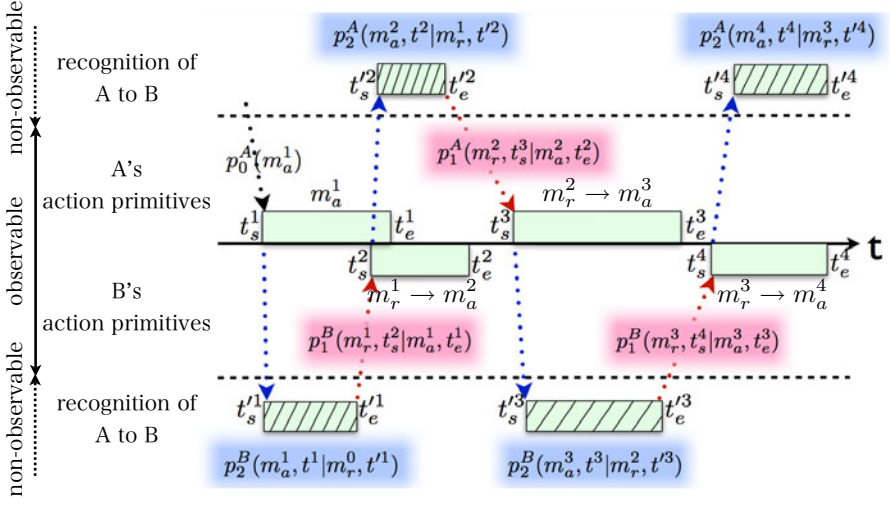


Fig. 1. Structure of the dialogue behavior model

- p_2 : the probability of transiting basic dialogue behaviors ($I_S^i \rightarrow I_S^{i+1}$)
 When the basic dialogue behavior I_S^i transits to I_S^{i+1} , the posterior motion primitive m_r^i by human B in I_S^i becomes the prior motion primitive m_r^{i+1} in the next basic behavior I_S^{i+1} . However, it does not always occur. This transition occurs only when human A needs to respond to a motion by human B . The probability p_2 of regarding the response m_r^i at the current step as the prior motion m_a^{i+1} at the next step is calculated as:

$$p_2^A(m_a^{i+1}, t^{i+1} | m_r^i, t^{i+1}) \quad (3)$$

where t^i denotes the recognition time against m_a^{i+1} . If the difference between the motion primitive by the asker and its recognition by the responder is large, the dialogue sometimes breaks. The probability seems to depend on invisible factors, such as the mental states, intentions, social status of each human, and the environment in which the dialogue occurs.

By using these probabilities, the observed dialogue behavior can be described by following a probability chain model:

$$p^A(m_a^1) p_2^B(m_a^1, t^1 | m_r^0, t^1) p_1^B(m_r^1, t_s^2 | m_a^1, t_e^1) p_2^A(m_a^2, t^2 | m_r^1, t^2) \cdots p_2^B(m_r^N, t^N | m_a^{N-1}, t^N) \quad (4)$$

The purpose of our study is to extract probability p_1 in the proposed model. By using the dialogue data while performing the same task as the one performed in the experiment for estimating p_2 , p_2 can be approximated to 1 except for the end time of motion primitive.

4 Extraction of Response Time Distribution

In this section, the method used to learn the transition probability of dialogue behaviors is described.

4.1 Extraction of Motion Primitive

In our study, body motions in dialogues are represented by motion primitives, which are the smallest units of motion. A motion primitive is similar to a phoneme used for phonetic recognition. Motion is described as a transition of motion primitives, i.e., typical postures. The advantage of our method is that it is less subject to individual differences in terms of physical features, because only the information obtained from postures is used to define motions. Another advantage is that the method does not require manual procedures, such as annotating postures. The method to extract motion primitives is described as follows.

Preprocessing

First, 3-D position data $\mathbf{x} = (x, y, z)^t$ of specific sites on the body of a speaker is measured at 30 fps by Microsoft's Kinect system using a range sensor and a camera. Eight positions on the body are measured: the head, the chest, both shoulders, both elbows, and both hands, as shown in Fig.2.

When the measurement of a 3-D position fails, the position is estimated by a linear interpolation between the previous point and the next point at the failure point. We used linear interpolation because the duration of failure were relatively short in the experiment and the movements of body site positions in the short term could be estimated by a uniform linear motion model.

The 3-D positions of the body sites in an absolute coordinate system were obtained using the Kinect system: however, it is necessary to convert coordinates in order to recognize that the postures are equivalent to the ones that have the same features but face in another direction or position, i.e., it is necessary to convert coordinates for posture recognition without depending on the direction and position of the posture. To overcome this problem, the coordinate origin is shifted to the center of the torso region and the body direction is normalized. This makes it possible to observe the relative motion as seen from the center of the torso region and to extract the motion primitive without depending on the human's position and direction.

Classification of Postures

The motion primitives are extracted from the obtained 3-D position data. First, joint angles of the upper body $\Theta = \{\theta_i | i = 1, 2, \dots, 8\}$ are obtained as a feature of the posture from the 3-D position data. The advantages of using joint angle information are as follows:

- The number of feature points is low.
- There is a poor correlation between each feature point.
- The posture recognition is less subject to individual differences.

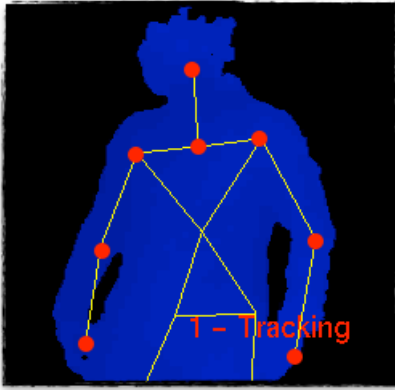


Fig. 2. The 3-D points of joints (red circles) measured by the Kinect system

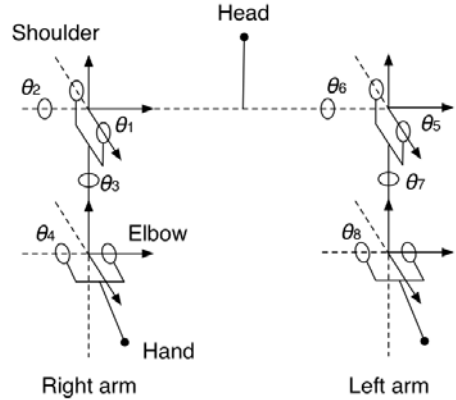


Fig. 3. Posture parameters

Each cluster of posture features generated by a clustering method based on a k -means algorithm is regarded as a motion primitive. The number of clusters is equivalent to the number of motion primitives and the postures in the same cluster are classified as the same motion primitive. Optimal k is estimated through experiments.

By this procedure, the transition of postures with joint angles are obtained as time-series data from each set of learning data.

4.2 Interval Detection of Utterances

Intervals of utterances are detected from speech waves obtained by a pin microphone attached to the speaker's clothes. In the detection method, the power of a speech wave is calculated by Fast Fourier Transform (FFT) with a divided speech wave and converted to a binary form using an adequate threshold. We interpolated using an adequate time threshold between two nearby intervals of utterances in order to deal with the case when the detection of an utterance interval fails owing to uneven voice volume.

4.3 Detection of Response Time Distribution

Based on the obtained information about motion primitive transitions, values for calculating the probabilities (p_1) of prior motion, posterior motion, and response time between prior motion and posterior motion were obtained. Response time is the time between the end of one (prior) motion and the start of next (posterior) motion. We extracted response times from a large amount of learning data and plotted histograms in order to estimate the shape of its probabilistic distribution. The accuracy of parametric estimation can be improved by iterative learning with estimated probabilistic distribution.

5 Experimental Results

5.1 Experimental Conditions

In this study, we conducted experiments to extract the response time distribution of interaction behaviors. We assumed the situation a wedding reception scenario and extracted the response time of dialogues and behaviors between a guest and a receptionist. In this situation, the start and end times of behavioral actions are clear, and many interactions such as offering greetings, hand-delivering wedding presents, and signing the guest book are performed at the reception. The behavioral data were collected by using the method described in the previous section and the layout of sensors is shown in Fig. 4. The red dashed line area in Fig. 4 shows the area where motion was captured by Kinect sensors. A guest book and a pen were placed on the table. The relative 3-D positions of joints and the time of utterance were stored as exemplars in a behavior database. The duration of an interaction at the reception area is about 1 min, and no restrictions are given to subjects except the scenario shown in Figs. 5, 6, 7 and 8.

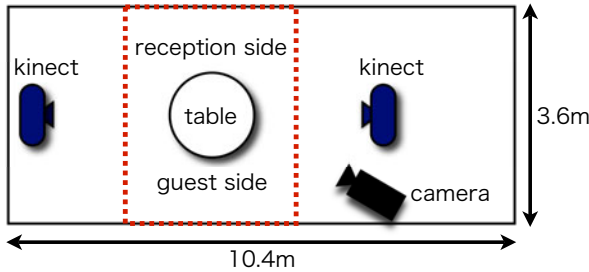


Fig. 4. Arrangement of the experimental environment

5.2 Extraction of Response Time Distribution

From the behavior database, we extracted motion primitives of behavior and the distribution of the response times between motion primitives. It is difficult to determine the number of motion primitives because if the number is too small we cannot capture the structure of the interaction, and if the number is too large we cannot extract meaningful distributions. In this study, we determined the number of action primitives to be six through the preliminary experiment. The extracted response time distribution is shown in Fig. 10 when the number of action primitives is six. Fig. 10 shows the histogram of each pair of motion primitives. As shown in Fig. 9, the horizontal axis shows the interval time for each 200 ms when pre-action primitive m^3 is completed and the vertical axis shows the frequency of the post-action primitive m^5 . A negative interval time means an overlap of action primitives.

When extracting the response time, we ignored the action primitives when the response time was smaller than -1200 ms or greater than 1400 ms, because the relationship between such action primitives is weak. The distribution shapes of the response times are divided into four types:



Fig. 5. Arrival of the guest



Fig. 6. Hand-delivery of presents



Fig. 7. Signing the guest book



Fig. 8. Guest leaves

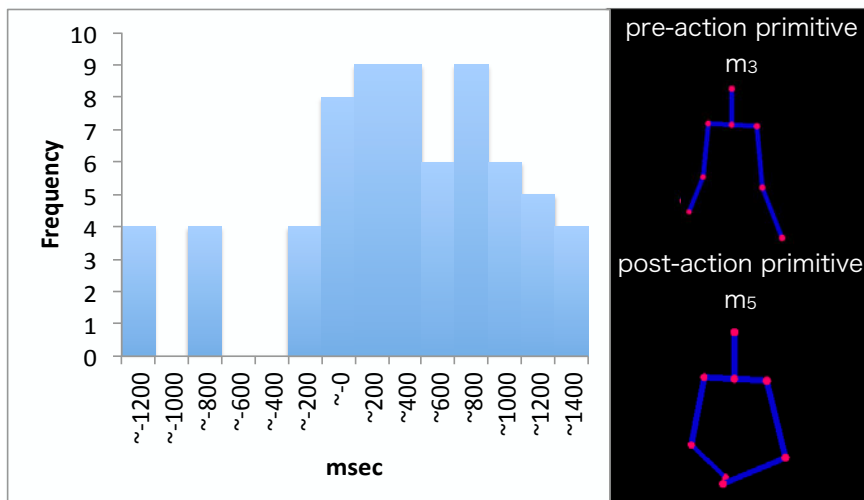


Fig. 9. The histogram of response time (pre-action primitive m^3 and post action primitive m^5)

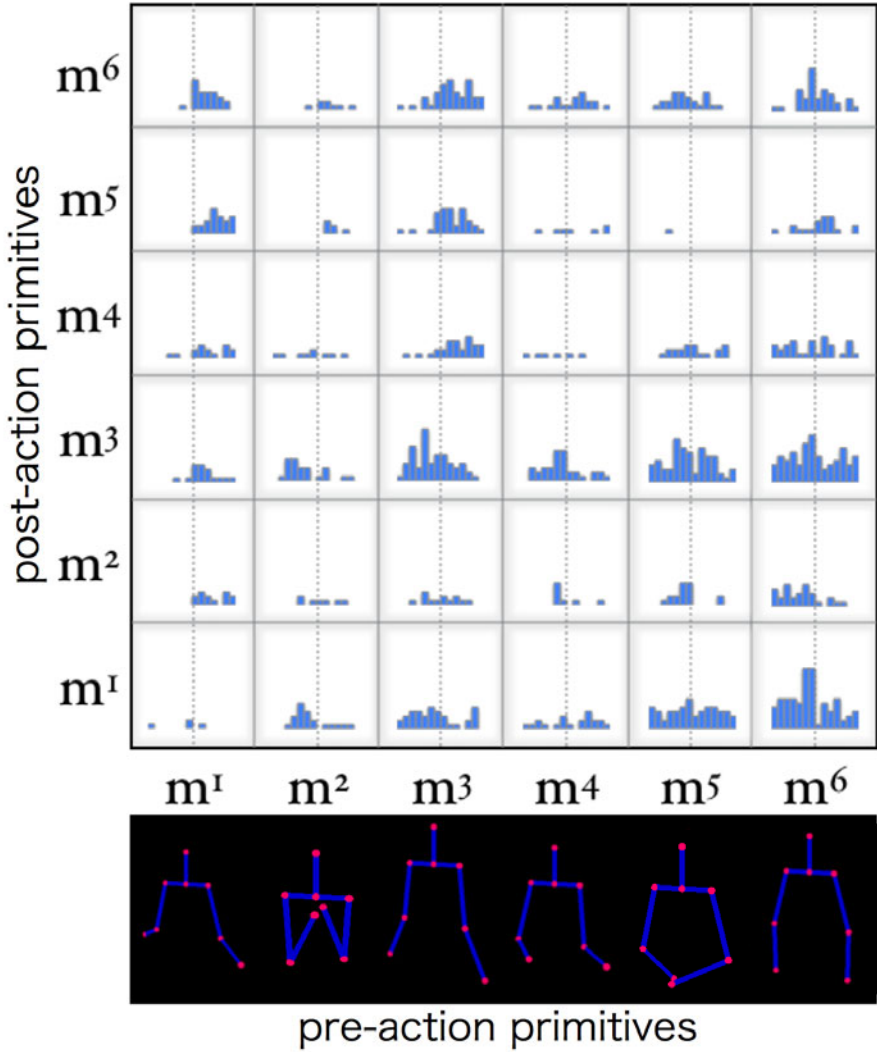


Fig. 10. Extraction of the response times when the number of action primitives is six

- (i) The distribution has a peak between the termination time of the preceding action and 1000 ms after the preceding action. (e.g., a pair of (m^3, m^6)),
- (ii) The distribution has a peak between -600 ms and the termination time of the preceding action (e.g., pairs of (m^5, m^3) and $(m^6 \text{ and } m^1)$),
- (iii) The distribution has no explicit peak (e.g., a pair of (m^5, m^1)),
- (iv) There is no correlation between action primitives (e.g., pairs of (m^2, m^4) and (m^4, m^5)).

In types (i) and (ii), the frequency of the following action primitive increases just after the preceding action primitives or when they overlap. There might be

a distribution of the response time with a peak same as that for an utterance. For type (iii), the results cannot be treated as a distribution with an explicit peak. The reasons of this phenomenon are lack of training data, in accuracy of action primitive extraction, or no correlation between action primitives in real time. In type (iv), the correlation between action primitives is very low, so the transition probability of such an interaction is very low.

5.3 Distribution Extraction under Utterance Effect

Next, we conducted an experiment to investigate the distribution changes affected by utterances. In this experiment, the post-action primitives are extracted only when an utterance is detected, and then compared with the results in Fig. 10. Figure 11 shows the distribution of the response time of the post-action primitives extracted only when an utterance is detected. Comparing Fig. 11 with Fig. 9, we observe that the peak of the distribution of the post action primitives is shifted forward. The reason for this shift is that the post-action primitives are suspended until the guest or receptionist finishes his or her utterance. From these results, it can be concluded that there exists a valid response time between the preceding action primitives and post action primitives, but the distribution of the response time changes when the preceding action primitives are performed with an utterance. Therefore, when activity support systems return a response, they are effective for creating visual and acoustic reply patterns.

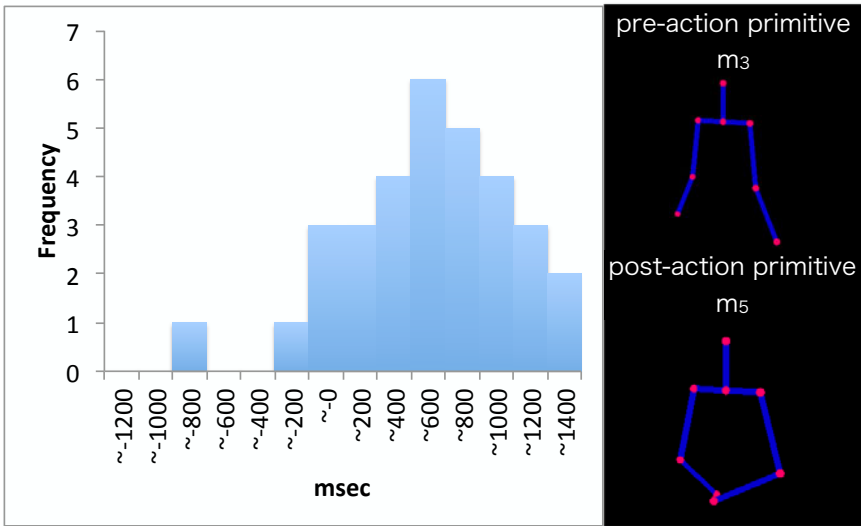


Fig. 11. The histogram of the response time when an utterance is detected (pre-action primitive m^3 and post-action primitive m^5)

6 Conclusion and Future work

In this study, we extracted the distribution of the response time of utterances and action primitives by detecting 3-D positions of joints and utterances. We also proposed a method to structurally model dialogues. In order to show the effectiveness of the response time between action primitives in a dialogue, we collected dialogue data from a wedding reception scenario and analyzed the distribution of the response times. From this analysis, we found that an appropriate interval existed for a response time of action primitives in the dialogue, and that the interval is affected and differs by a pair of action primitives and the presence of an utterance.

In future work, we will investigate a method for recognizing multi modal dialogue by using structural representation of human interactions while maintaining an appropriate interval between action primitives. This will lead to the realization of a system that can enable dialogue between users, which appear more natural, by using the structural representation of action primitives for generating responses from the system.

Acknowledgement. This work is partially supported by cooperative research with Daiwa House Industry Co., Ltd. and by Grant-in-Aid for Scientific Research on Innovative Areas (No. 22118506).

References

1. Takizawa, M., Makihara, Y., Shimada, N., Miura, J., Shirai, Y.: A Service Robot with Interactive Vision-Object Recognition Using Dialog with User. In: Proc. of the 1st Int. Workshop on Language Understanding and Agents for Real World Interaction (Academic Journal), pp. 16–23 (July)
2. Kuriyama, H., Murata, Y., Shibata, N., Yasumoto, K., Ito, M.: Congestion Alleviation Scheduling Technique for Car Drivers Based on Prediction of Future Congestion on Roads and Spots. In: Proc. of 10th IEEE Int'l. Conf. on Intelligent Transportation Systems (ITSC 2007), pp. 910–915 (September 2007)
3. Fukaya, K., Watanabe, A.: Intuitive Manipulation to Mobile Robot by Hand Gesture. In: 24th ISPE International Conference on CAD/CAM, Robotics and Factories of the Future (July 2008)
4. Shirberg, E.: Spontaneous Speech: How People Really Talk and Why Engineers Should Care. In: Proc. EUROSPEECH (2005)
5. Fujie, S., Fukushima, K., Kobayashi, T.: Back-channel Feedback Generation Using Linguistic and Nonlinguistic Information and Its Application to Spoken Dialogue System. In: Proc. EUROSPEECH, pp. 889–892 (2005)
6. Hirose, K., Sato, K., Minematsu, N.: Emotional speech synthesis with corpus-based generation of F0 contours using generation process model. In: Proceedings of International Conference on Speech Prosody, Nara, pp. 417–420 (March 2004)
7. Fujiwara, N., Itoh, T., Araki, K.: Analysis of Changes in Dialogue Rhythm Due to Dialogue Acts in Task-Oriented Dialogues. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 564–573. Springer, Heidelberg (2007)

8. Nishimura, R., Kitaoka, N., Nakagawa, S.: Analysis of relationship between impression of human to human conversations and prosodic change and its modeling. In: Proceeding of the Interspeech, pp. 534–537 (2008)
9. Kawashima, H., Nishikawa, T., Matsuyama, R.: Analysis of Visual Timing Structure in *Rakugo* Turn-taking (written in Japanese). *IPSJ Journal* 48(12), 3715–3728 (2007)
10. Sumi, Y., Yano, M., Nishida, T.: Analysis environment of conversational structure with nonverbal multimodal data. In: 12th International Conference on Multimodal Interfaces and 7th Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI 2010), Beijing, China (November 2010)

A Flexible Method for Localisation and Classification of Footprints of Small Species

Haokun Geng¹, James Russell², Bok-Suk Shin¹,
Radu Nicolescu¹, and Reinhard Klette¹

¹ Department of Computer Science, University of Auckland, Auckland, New Zealand
hgen001@aucklanduni.ac.nz, {b.shin,r.nicolescu,r.klette}@auckland.ac.nz

² School of Biological Sciences, Department of Statistics, University of Auckland,
Auckland, New Zealand
j.russell@auckland.ac.nz

Abstract. In environmental surveillance, ecology experts use a standard tracking tunnel system to acquire tracks or footprints of small animals, so that they can measure the presence of any selected animals or detect threatened species based on the manual analysis of gathered tracks. Unfortunately, distinguishing morphologically similar species through analysing their footprints is extremely difficult, and even very experienced experts find it hard to provide reliable results on footprint identification. This expensive task also requires a great amount of efforts on observation. In recent years, image processing technology has become a model example for applying computer science technology to many other study areas or industries, in order to improve accuracy, productivity, and reliability. In this paper, we propose a method based on image processing technology, it firstly detects significant interest points from input tracking card images. Secondly, it filters irrelevant interest points in order to extract regions of interest. Thirdly, it gathers useful information of footprint geometric features, such as angles, areas, distance, and so on. These geometric features can be generally found in footprints of small species. Analysing the detected features statistically can certainly provide strong proof of footprint localization and classification results. We also present experimental results on extracted footprints by the proposed method. With appropriate developments or modifications, this method has great potential for applying automated identification to any species.

1 Introduction

Computer-based systems have been a common technique of humankind to perform activities that have to be repeated numerous times [5]. Identifying small species from their footprints is one of such activities. Currently ecological experts need to spend much effort and time identifying footprints from inked tracking cards, highly dependent on the experts' knowledge and experiences, and the manual identification analysis often needs to be repeated on many tracking cards.

Therefore, we would like to initialise a work based on many previously researched theoretical findings, with knowledge from ecology, especially from the

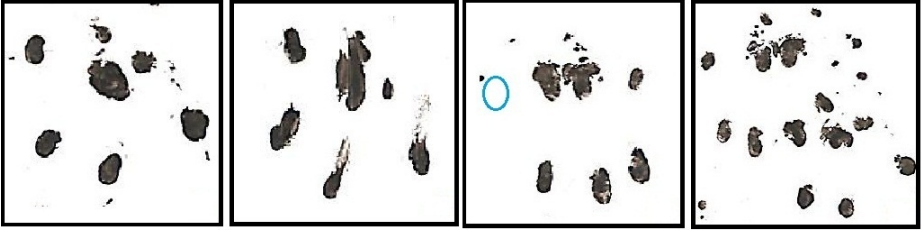


Fig. 1. Samples of mice footprints in different situations. From left to right: Normal front footprint, sliding front footprint, missing toe hind footprint, and overlapped hind&front footprints

study area of track recognition, and to transfer them into a practical technique to assist ecological experts in analysing inked tracking cards.

The demand for such systems that can process the automated identification of species from their scanned footprint images is most likely to increase in the future [8]. It becomes essential to have a working application that can be properly incorporated into the current system, handles the repetitive jobs, and outputs accurate and reliable results.

However, the presentations of footprints are varied, the ‘puzzle’ is that the images of a footprint may have very different appearances (as shown in *Fig.1*). Besides normal footprints, other undesirable image data include sliding footprints, missing toe footprints, and overlapped footprints. A single tracking card may contain footprints from ≥ 1 individuals.

Before any further analysis can be carried out by the automated recognition algorithm, those varied representations of the footprints need to be transformed into digitalised geometric models. Correctly understanding and handling the transformation process is certainly a difficult task.

In this paper, we will firstly introduce the current standard track acquisition procedure. Then we will describe how a footprint can be digitalised and understood by our image processing application. This involves a rule based footprint recognition algorithm that performs automated footprint localisation and classification. Additionally, we would like to present the conceptualization of an integrated proposed system, and suggest some possible future work.

2 Track Acquisition of Small Species

The *Tracking Tunnel System* is a widely used standard procedure for collecting tracks of small animals to gain an index of the abundance of target small species in New Zealand [2]. It is a cost-effective method to collect tracks of small species over large areas [10]. Providing reasonable analysis and reliable results on the estimate of species’ presence plays an important role in ecological research when ecologists decide to study rare species or assess community composition for environmental surveillance or pest control [7].

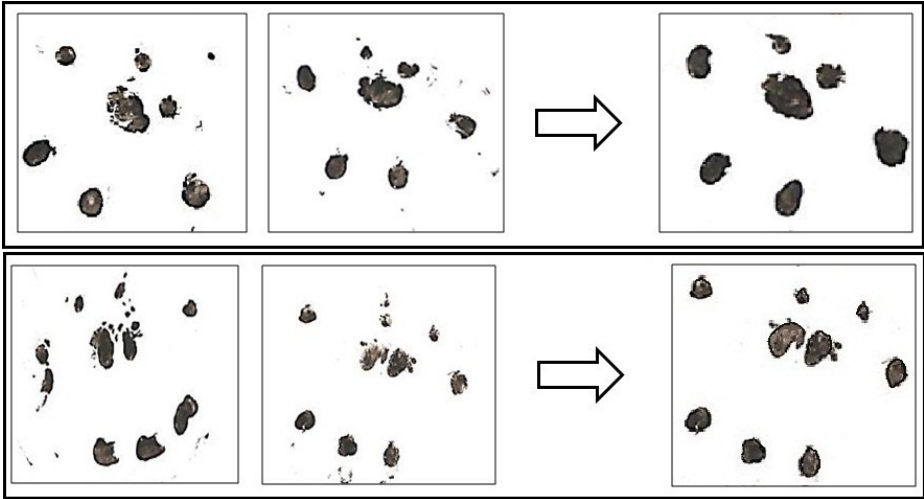


Fig. 2. *Top:* Isolated front footprint model example. *Bottom:* Isolated hind footprint model example.

Traditionally, tracks or footprints are collected by this tracking tunnel system, and the identification of tracks and footprints is handled manually by experienced wildlife experts [3]. The basic principle of animal tracking is firstly to recognise single footprints from a number of unknown footprints, and then to identify the species based on the analysis of its footprints [10].

The tracking tunnel system is considered the first step when ecologists would like to non-invasively monitor or study a selected species. The collected tracks or footprints need to be analysed manually by human experts. In the identification procedure, distinguishing among many morphologically similar species through analysing their footprints is extremely difficult, and one single tracking card can also contain footprints from different species [10]. Our method aims to ultimately implement an automated recognition process to assist experts in the current identification procedure.

3 Footprint Geometric Analysis

The further implementation of the track recognition algorithm would highly depend on the understanding of the footprint geometric models of targeted species. In the following experiments, we choose house mice (*Mus musculus*) as our major object of study.

We isolate normal footprints from many tracks on a tracking card. The front foot for a house mouse usually has four toes, the hind foot usually has five toes [10]. The toes of the front foot are evenly distributed around the central pad. The hind foot normally has three toes bunched in front of the central pad that can roughly form a straight line. Based on the previous studies [10] and our experimental set of tracking cards, we isolated normal footprints from tracks.

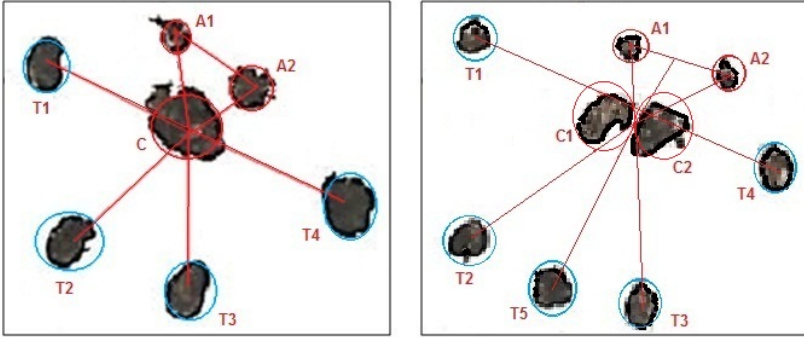


Fig. 3. *Left:* Geometric model for mice front footprint. *Right:* Geometric model for mice hind footprint.

The isolated front footprint model is shown in *Figure 2* (top), and the isolated hind footprint model is shown in *Figure 2* (bottom).

Manually analysing the relations among toes and central pads of the isolated footprints is an essential step for footprint localisation and classification (i.e., front or hind footprints). *Figure 3* visually represents how geometric features will be analysed by the proposed algorithm.

Figure 3 (left) is an isolated front footprint, it has a clear geometric structure: a central pad is in the centre of the footprint. There are two accessory pads and four toes evenly distributed around the central pad. In most cases, the central pad has the larger size than other nodes in this particular region. The toe prints are marked by blue circles, the central pad and accessory pads prints are marked by red circles. The central pad is distributed in the middle point of that line segment. The central pad and two accessory pads clearly form a triangle. Also there are three straight lines intersecting the central pad, they are $\overline{T_1T_4}$, $\overline{T_2A_2}$, and $\overline{T_3A_1}$.

Figure 3 (right) is an isolated hind footprint, it has a similar geometric structure to the front footprint. However, by contrast it has a differently formed central pad which consists of two vice pads. It also has three toes in the front, they can roughly form a straight line $\overline{T_2T_3}$ that is approximately parallel with the line formed by the two outer toes $\overline{T_1T_4}$. Comparing with the front footprint, the formation of the hind central pads and the number of toes could be two significant conditions of front and hind footprints classification.

4 Footprint Extraction

We propose a method for extracting regions of interest using improved *OpenSurf* libraries¹ with the rule-based conditional filtering and geometric model. This

¹ The term ‘‘Open’’ refers to one of its major development components *OpenCV*, and ‘‘SURF’’ is the abbreviation of *Speeded Up Robust Features*.

method detects significant interest points from many distributed points on input images, and it extracts regions of interest which are suitable for geometric model analysis.

4.1 Interest Points Detection

The implementation of OpenSURF libraries was based on an interest point detection-description scheme, first described by Bay et al. [1] in 2006. An intermediate image representation plays an important role in the improvement of SURF's performance, which is known as the "Integral Image" [11]. The integral image can be computed directly from the input image by the following formula:

$$I_{\Sigma}(x, y) = \sum_{i=0}^x \sum_{j=0}^y I(x, y)$$

where I is the input image, and (x, y) are the x- and y-coordinates of a certain pixel on the input image. So the integral image I_{Σ} can be then calculated by the formula given above [4].

Figure 4 shows a sample scanned tracking card image with many interest points detected before applying the filtering function. The blue circles indicate the interest points that have black marks on a white background with a detected radius.

However, it is too difficult to define the proper regions of interest on the input image at this stage, because the representation of small species' footprints

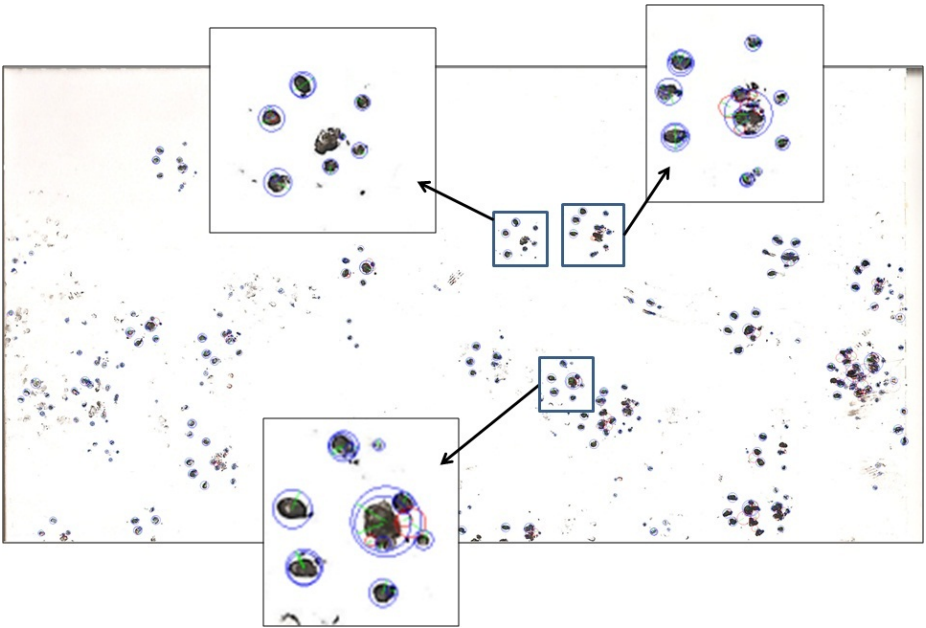


Fig. 4. Raw analysis result for a sample tracking card image

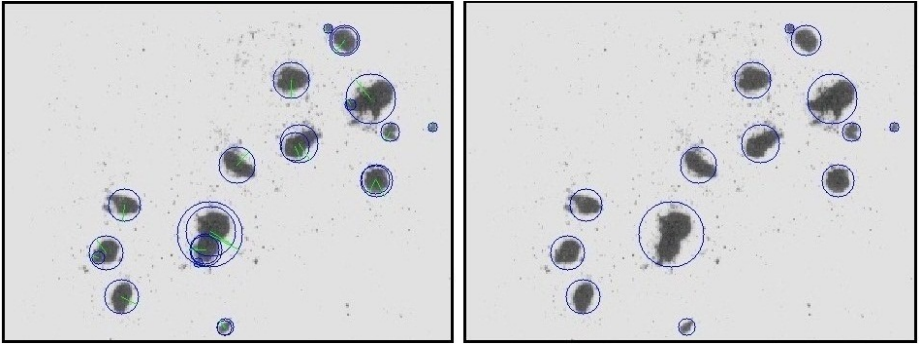


Fig. 5. *Left:* Detected interest points before preprocessing. *Right:* Identified areas of interest after preprocessing.

are commonly massively distributed, and too many interest points are detected, including many insignificant points. This certainly requires a filtering function to remove all irrelevant interest points. Therefore we implemented a rule-based filtering algorithm to remove insignificant interest points.

4.2 Regions of Interest Extraction

Applying this rule-based conditional filtering function, most of the insignificant or noisy interest points can be detected and removed. After applying a rule-based conditional filtering function, most of the insignificant or noisy interest points can be detected and removed. *Figure 5* (Left) shows all the initially detected interest points on an input image. *Figure 5* (Right) presents the experimental result after using filtering rules. Basically all the regions of interest on the input image are detected and localised correctly. For the filtering algorithm, see *Figure 7*.

Analysing the standard geometric models could provide us with the following organised truths of our study object's footprints:

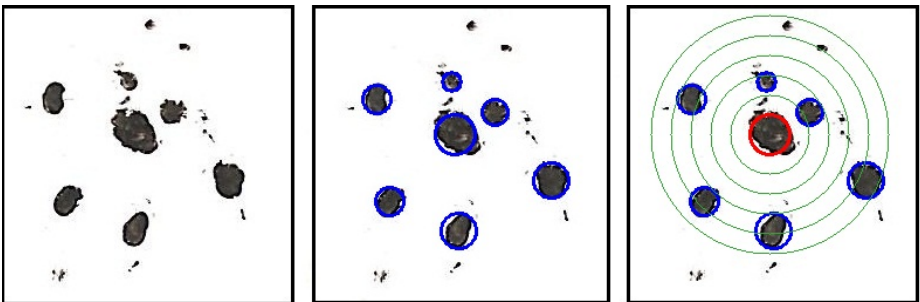


Fig. 6. *Left:* Original image. *Middle:* Interest points preprocessed by our application. *Right:* Central pad localisation by our application. Blue circles indicate 'area of interest'; red circle indicate 'recognised central pad'; green circles indicate distance from the centre of the central pad, each gap represents one times the central pad radius.

```

1: Initialise L; {the list of interest points on the input image}
2: for (each interest point I in list L) do
3:   if (the radius of I is less than 6 pixels) then
4:     Delete the interest point from list L;
5:   else
6:     Initialise V; {an empty interest point list }
7:     for (each interest K in the list L) do
8:       if (the radius of K  $\geq$  the radius of I) then
9:         Initialise D; { the distance between I and K }
10:        if (the radius of K  $\geq$  the radius of I + D ) then
11:          if (the radius of K  $\geq$  7/10 of the radius of I) then
12:            Store K in V
13:          else
14:            Store I in V
15:          end if
16:        else
17:          if (D < 6) then
18:            if (the radius of K  $\geq$  7/10 of the radius of I) then
19:              Store K in V
20:            else
21:              Store I in V
22:            end if
23:          else
24:            Store I in V
25:          end if
26:        end if
27:      end if
28:    end for
29:    for (each interest point T in V) do
30:      Delete T from the list L;
31:    end for
32:  end if
33: end for

```

Fig. 7. Pseudo-code for the pre-processing stage

Accurate matching: a central pad normally has the largest area within the six times its radius bounded region, and there should be exactly six (for the front footprint) or seven (for the hind footprint) smaller areas of interest in that particular region.

Loose matching: a central pad normally has a radius larger than the average radius within the six times its radius bounded region, and the number of areas of interest in this particular region should be greater than or equal to four, and less than or equal to ten.

Figure 6 shows the progress of locating a possible central pad on an input image. We use green circles to indicate the distance from the centre of the central pad, which also is the centre of the possible region for a footprint. From the inside to the outside boundary, each gap between every two green circles represents the

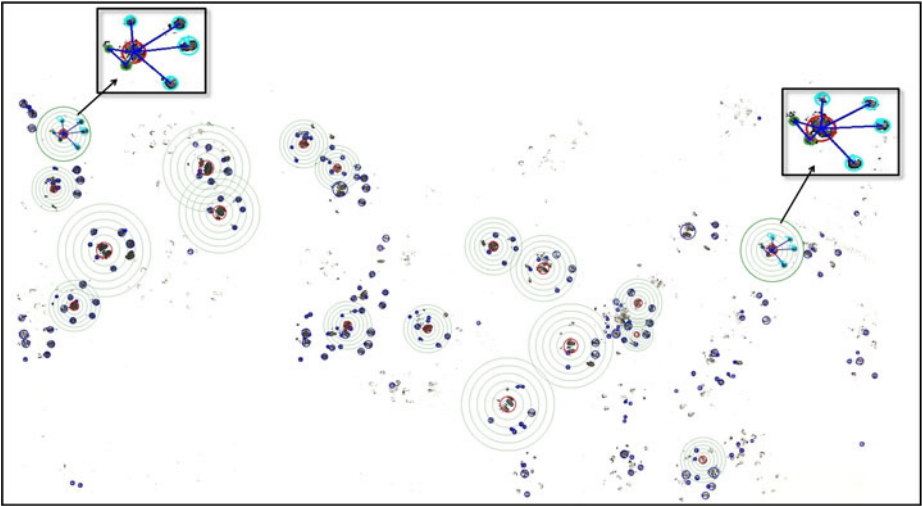


Fig. 8. Footprint detection result of an input tracking card image. Fully recognised footprints are indicated by zooming in.

length of the radius of the central pad area. The outside boundary shows the region of a possible footprint on the image. This region could be valuable when human experts decide to do manual additional analysis of the tracking card.

As the central pad can be recognised, the region of a possible footprint is located with a proper boundary, which is six times the radius of the central pad. The algorithm can then test the interest points within this range for whether their distribution matches the pre-defined model. We again defined a rule-based approach for the footprint identification and localisation. *Figure 8* shows the result image after processed by the following rules:

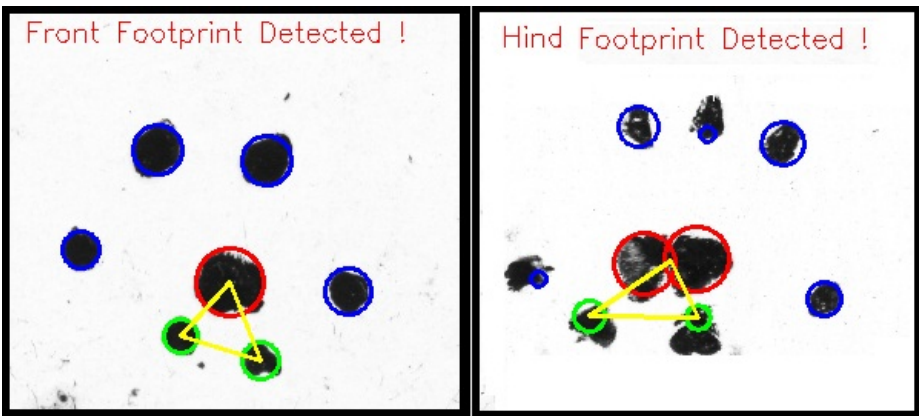


Fig. 9. Testing results for front and hind footprints

- Rule 1.** Two accessorial pads should be close to the central pad, generally within the range of three times the radius of the central pad.
- Rule 2.** Two accessorial pads must have smaller distance to each other than their distance to other areas of interest in this particular region.
- Rule 3.** Two accessorial pads and the central pad can form a triangle at the back of the footprint. Each angle inside the triangle should be only smaller than or equal to 90° , and the sum of the three angles is exactly 180° .
- Rule 4.** A line segment can be drawn between every two toes. The longest line segment, which is the line between the left and the right outer toes, must cross the area of the central pad.
- Rule 5.** If the region with a recognised central pad can not completely match all the rules, it should be considered and marked as a ‘possible region of a footprint’ on the result image for human experts to review.
- Rule 6.** If the region with a recognised central pad has more than ten areas of interest within its considerable range, which is six times the radius of the central pad, the algorithm should identify this region as an ‘unpredictable region’, and it will refuse to do any further analysis.
- Rule 7.** If only one central pad is detected in this region, and also two accessorial pads and four toes are detected, the algorithm will identify it as a front footprint of our study object.
- Rule 8.** If two central pads are detected in this region, and also two accessorial pads and five toes are detected, the algorithm will identify it as a hind footprint of our study object.

Since normally every two nearby toes have certain angles in between, a statistical analysis was used to find out the angles between every two nodes of the footprint samples. The corresponding statistical analysis of those angles provides us the following additional classification rules:

Front or hind footprint classification: for front footprints, the average value for angle $\angle T_2CT_3$ is 46.2° in the range from 43.6° to 48.9° ; for hind footprints, the average value for angle $\angle T_2CT_3$ is 56.1° in the range from 53.2° to 59.8° . There is a clear difference between the two ranges.

Left or right footprint classification: if angle $\angle A_1CT_1$ is less than angle $\angle T_4CA_2$, then this is a left footprint; otherwise, this is a right footprint.

Figure 9 presents two fully recognised footprints of our study object with expected classification results. In this case, the algorithm counts the number of central pads detected, a front footprint should have one central pad whereas a hind footprint should have two. Also the number of toes are different, a front footprint should have four toes, and a hind footprint should have five.

4.3 Geometric Feature Extraction

By understanding the standard geometric models of our study object’s footprints, we could gather useful information for relations among nodes within one region of interest, such as angles, areas, or distances. These geometric features can be generally found by footprints which are left by small species. Analysing

the detected features statistically can certainly provide strong proof of footprint localisation and classification results. We extract geometric features from associated nodes in the regions of interest. Data collection is only as valid as the feature extraction, and is considered an important objective of our application.

In order to collect statistical analysis data for classifying similar species (e.g. rats and mice), we have to sort the accessorial pads and toes in a certain order (e.g. sort them clockwise). The application would collect (1) radius of each node; (2) distance between each node and central pad; (3) the area of the triangle (formed by central pad and its accessorial pads); (4) internal angles of that triangle, etc. These statistic data will be used to find definitive differences for distinguishing morphologically similar species, such as rats and mice.

4.4 Directional Scale Vector

Every footprint has a relative direction. The footprint directions are indispensable when we try to find a single track of an individual mammal. In our approach, we use a 2D vector \mathbf{v}_{ds} to present the relative direction of a given footprint, it is called *directional scale*. Based on the extracted geometric regions and their centroids, the directional scale vector $\mathbf{v}_{ds} = (x_s, y_s)$ can be calculated by the following formula:

$$x_s = \left(\sum_{i=1}^{n_r} x_i \right) - n_r \cdot x_{cp} \quad \text{and} \quad y_s = \left(\sum_{i=1}^{n_r} y_i \right) - n_r \cdot y_{cp}$$

Here, n_r is the number of regions (i.e. toes and accessorial pads of the given footprint, not counting the central pad), x_{cp} and y_{cp} are the x - and y -coordinates of the centroid of the central pad, and x_i and y_i are the coordinates of the centroid of region i . This vector can precisely indicate the relative directions of any footprints. By picking up those footprints with similar direction scale values, we can then directly indicate which path on the tracking card the current footprint belongs to.

5 Experimental Test

In order to systematically test the accuracy of our algorithm, we test accuracy on two datasets of mice tracked on cards. The first study is of introduced house mice from New Zealand and the second study of hazel dormice from the United Kingdom. The combined image data set has been divided into three data groups (as shown in *Figure 10*). We recorded the experimental data during application of the track recognition algorithm. The experimental results are reported by tracking card for the three different image data groups in four different categories, where each category represents a classification related to regions of interest on a card. Thus a single card with multiple regions of interest may fall in more than one of the following categories explained below, depending on the number of regions identified:

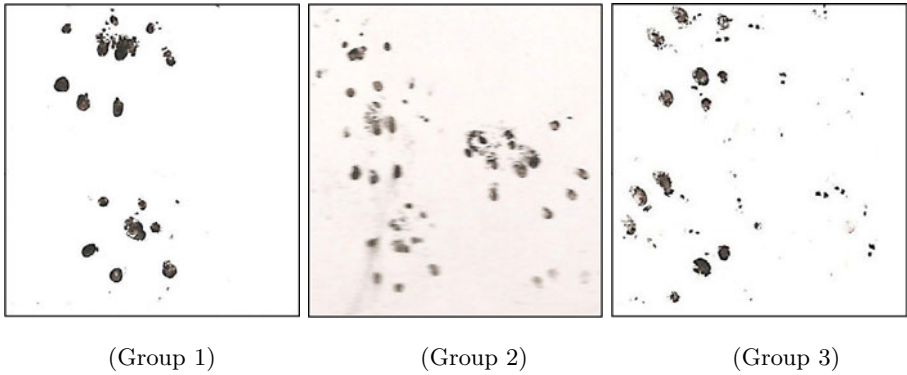


Fig. 10. The preview images of the three image data groups. (Group 1) clear foreground and background; (Group 2) dim foreground and background; (Group 3) unexpected species involved (here an invertebrate).

Sensitive Matches: The number of cards containing an identified footprint fully matching the pre-defined rules in the algorithm (i.e. confirmed species presence).

Loose matches (True): The number of cards containing a region of interest loosely identified as a possible region of footprints, where it is truly one footprint (i.e. footprint identified).

Loose matches (False): The number of cards containing a region of interest loosely identified as a possible region of footprints, but it is not a footprint (i.e. mis-identification).

Did not detect print: The number of cards where no footprints are detected by the algorithm (i.e. species absence).

The experimental result (as shown in Table 1) indicates that the algorithm has a fairly high success rate for sensitive footprint identification and loose-condition matches for images with clear prints and clean background.

The accuracy for dim background and foreground images is lower than the results for images from clear background and foreground images. In addition, the accuracy for tracking cards with tracks from unexpected species

Table 1. The experimental statistical analysis for the algorithm accuracy evaluation. “Sensitive matches” indicates rate of best matched footprints. Correctly or incorrectly detected possible footprints are assigned as “loose matches (true)” or “loose matches (false)”. “Did not detect print” records no footprints detected for a card.

Classification	Percentage of Detection Accuracy		
	Group 1 (72 cards)	Group 2 (42 cards)	Group 3 (22 cards)
Sensitive matches	77.8 %	61.9 %	68.2 %
Loose matches (True)	85.7 %	68.3 %	80.7 %
Loose matches (False)	14.3 %	31.7 %	19.3 %
Did not detect print	1.4 %	9.5 %	9.1 %

(e.g. invertebrates) is surprisingly good; the reason being that the track recognition algorithm has a filtering function that filters out all the tracks with very small regions, which might be left by unknown species other than our object of study (e.g. invertebrate tracks).

6 Conclusions

We propose a method for locating and classifying footprints of small species on scanned tracking card images by three major steps: (1) extracting regions of interest; (2) further analysis with “rule-based conditional filtering”; (3) extracting footprint “geometric features”. In addition, our method can provide useful results for finding footprints which belong to the same path with a 2D vector called “direction scale”.

Comparing with some previous studies [6,8,9,10,12] in this research field, we propose two new ideas for this algorithm:

A footprint could be identified either “fully matched” or “partially matched”, which depends on the degree of matching the pre-defined rules. Due to the sparse amount of information provided by the detected interest points, rule-based identification processes could be a key to the shortage of information. Moreover, rule-based identification could allow developers to add a new rule or modify the existing rules. This provides a greater extensibility to this algorithm.

Footprint geometric models could provide precise mathematical relationships among nodes of the standard footprint for any target species. In practical implementations, numbers, equations and formulas are always considered useful information for footprint identification (e.g. [10]).

The experimental results provide positive feedback on the accuracy of this algorithm; if the image cards have clear prints and clean background, 85.7 % of them can be detected as “loosely matched” by the pre-defined rules of this algorithm.

Acknowledgements. We are grateful to Jamie MacKay from the University of Auckland and Cheryl Mills from the University of Exeter for access to tracking cards ($n = 50$ and $n = 86$ cards, respectively).

References

1. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
2. Blackwell, G.L., Potter, M.A., McLennan, J.A.: Rodent density indices from tracking tunnels, snap-traps and Fenn traps: do they tell the same story? *New Zealand Journal of Ecology* 26, 43–51 (2002)
3. Brown, K., Moller, H., Innes, J., Alterio, N.: Calibration of tunnel tracking rates to estimate relative abundances of ship rats and mice in a New Zealand forest. *New Zealand Journal of Ecology* 20, 271–275 (1996)

4. Evans, C.: Notes on the OpenSURF library. University of Bristol (January 2009), www.cs.bris.ac.uk/Publications/Papers/2000970.pdf (last accessed on July 15, 2011)
5. Gaston, K.J., O'Neill, M.A.: Automated species identification: why not? *Philosophical Transactions of the Royal Society B* 359, 655–667 (2004)
6. Lowdon, I.M.R., Seaber, A.V., Urbanlak, J.R.: An improved method of recording rat tracks for measurement of the sciatic functional index of deMedinaceli. *Journal of Neuroscience Methods* 24, 279–281 (1988)
7. MacKenzie, D.I., Nichols, J.D., Lachman, G.B., Droege, S., Royle, J.A., Langtimm, C.A.: Estimating site occupancy rates when detection probabilities are less than one. *Ecology* 83, 2248–2255 (2002)
8. Mayo, M., Watson, A.T.: Automatic species identification of live moths. *Knowledge-Based Systems* 20, 195–202 (2007)
9. Medinaceli, L.D., Freed, W.J., Wyatt, R.J.: An index of the functional condition of rat sciatic nerve based on measurements made from walking tracks. *Experimental Neurology* 77, 634–643 (1982)
10. Russell, J.C., Hasler, N., Klette, R., Rosenhahn, B.: Automatic track recognition of footprints for identifying cryptic species. *Ecology* 90, 2007–2013 (2009)
11. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: *Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 511–518 (2001)
12. Watts, C.H., Thornburrow, D., Green, C.J., Agnew, W.R.: Tracking tunnels: a novel method for detecting a threatened New Zealand giant Weta. *New Zealand Journal of Ecology* 32, 92–97 (2008)

Learning and Regularizing Motion Models for Enhancing Particle Filter-Based Target Tracking

Francisco Madrigal, Mariano Rivera, and Jean-Bernard Hayet

Centro de Investigación en Matemáticas,
Jalisco s/n, Col. San Javier
36240 Guanajuato, GTO, México

Abstract. This paper describes an original strategy for using a data-driven probabilistic motion model into particle filter-based target tracking on video streams. Such a model is based on the local motion observed by the camera during a learning phase. Given that the initial, empirical distribution may be incomplete and noisy, we regularize it in a second phase. The hybrid discrete-continuous probabilistic motion model learned this way is then used as a sampling distribution in a particle filter framework for target tracking. We present promising results for this approach in some common datasets used as benchmarks for visual surveillance tracking algorithms.

1 Introduction

Visual target tracking has been the object of a very large amount of research in the last two decades, mainly in the communities of computer vision, image processing and networks. This is motivated in particular by a strong demand of automatic tracking tools for applications such as video-conferencing, gesture analysis, TV broadcasting, wildlife studies or video-surveillance, among many others. Our work targets particularly tracking in video-surveillance applications, which is characterized by a number of specific problems. First, tracking in that case is in general limited to pedestrians or cars. Moreover, a scene monitored by a surveillance camera typically contains many potential targets to track. They are not known in advance and have to be detected automatically. Furthermore, they generally undergo partial to complete occlusions, both from scene clutter (walls, pillars, poles. . .) or from other targets. Last, it is common in outdoors scenes that the appearance of different targets is quite similar, i.e. it may be hard to distinguish one target from another. In this difficult context and when dealing with only one camera, the motion model, i.e. the a priori knowledge about how objects move in the scene, helps to keep track of targets that are either ambiguous (if their appearance is similar to the one of a neighbour target) or occluded. The most simple and common motion models in the literature are constant velocity or constant acceleration ones, but they may not handle the specificities of a particular scenario: For example, because of the topology of one place, it may be frequent that pedestrians make sharp turns, which cannot be handled by simple motion models. Our contribution has been to infer a more complex probabilistic

motion model, by using the low-level information collected by the camera. To the best of our knowledge, this approach has not been proposed before, and the results we present in Section 5 show how promising it is. We will detail the construction of the prior in Section 3 and see how to use it in a tracking algorithm in Section 4. After presenting experimental results, we make a conclusion and give hints for future work in Section 6.

2 Related Work

Much progress has been done recently in the literature for the design of efficient tracking techniques. If early works have essentially imported filtering techniques from the radar community (e.g., the Kalman filter), later works have heavily relied on probabilistic models for the target appearance, e.g. Meanshift and its variants [5], or color histogram-based models incorporated into particle filters [13]. The latest trend in the area of multiple target tracking is the paradigm of tracking-by-detection [1,2], i.e. the coupling of tracking with powerful machine learning-based object detection techniques, that detect, by parts or as a whole, objects from some given class of interest (in particular, pedestrians [4]) with high confidence levels. The problem of tracking is then transformed into an optimization problem that finds a partition of space-time detections into coherent trajectories. The main limitation of these approaches is that their success depends essentially in the pedestrian detector. If it fails, the whole tracking system becomes inefficient. The problem is that in presence of occlusions, as it occurs frequently in our case, pedestrian detectors are likely to fail. Other detection-free techniques have been proposed [3] more recently, but they share with detection-based methods the need to wait for a given window of time before completing the association problem.

Here, we use a more traditional approach for tracking, namely the particle filter. We do not focus on the observation model, but rather on the other key element of any probabilistic tracker, the motion model. Works related to ours, i.e. based on developing new probabilistic motion models, is for example the one of [6], where the authors use a database of 3D motions to design a probabilistic model for tracking articulated objects. In this work, we aim at capturing the complexity of 2D motion in a given scene into a probability distribution. Such a complexity may be the consequence of specific physical elements (e.g., walls, corridors...) and it results in simple motion models being incorrect.

3 Learning Motion Transition Models from Sequences

This section describes our framework to learn motion priors from the low level information extracted from video sequences.

3.1 Estimating Empirical State Transition Priors

The idea of our approach is to learn information on state transitions, where the state refers to the quantities that will be estimated during tracking, in our case

position and velocity. We will denote these quantities as $\mathbf{r} = (x, y)^T$ (position) and $\mathbf{v} = (v^m, v^\theta)^T$ (velocity, decomposed in a magnitude v^m and an orientation v^θ). From optical flow information computed in video sequences taken by the same camera that will do the tracking, we collect statistics about the *temporal* evolution of this state. It is done as compactly as possible, to make the handling of these distributions tractable. For example, if we refer to the time as t , we are interested in learning information about the joint distribution $p(\mathbf{v}_{t+1}, \mathbf{r}_{t+1} | \mathbf{v}_t, \mathbf{r}_t)$, that sums up the a priori knowledge about where and at what velocity targets tend to be at time $t + 1$, when their position and velocity at t are given. Modeling this joint distribution in the continuous domain would require parametric distributions (e.g. mixtures of Gaussian distributions) difficult to handle, e.g. in a regularization framework. Similarly, a fully discretized distribution would require a lot of memory. Hence, we adopted a hybrid representation, partly discrete, partly continuous. We factorize the probabilistic state transition model as follows, by supposing conditional independence between the two velocity components, given the previous state,

$$p(\mathbf{v}_{t+1}, \mathbf{r}_{t+1} | \mathbf{v}_t, \mathbf{r}_t) \approx p(v_{t+1}^m | \mathbf{v}_t, \mathbf{r}_t) p(v_{t+1}^\theta, \mathbf{r}_{t+1} | \mathbf{v}_t, \mathbf{r}_t). \quad (1)$$

For the first term, we adopt a simple, continuous Gaussian model, i.e. $v_{t+1}^m \sim \mathcal{N}(v_t^m, \sigma_m^2(\mathbf{v}_t, \mathbf{r}))$. For the second term, which is a priori more complex in nature, as it represents potential direction changes, we use a discrete distribution. We get estimates for these two distributions from optical flow data.

Collecting low-level information about state transitions. Our basic low-level data are point tracks given by a sparse optical flow algorithm. In our implementation, we used the Lucas-Kanade (LK) algorithm [7]. Note that this algorithm keeps track of only well-defined image points, i.e. points that are locally non-ambiguous, and that have two large eigenvalues in their autocorrelation matrix [8]. This makes the collected data a priori reliable as we do not take into account textureless areas or edge areas (in which the aperture effect applies); this would not be possible with a *dense* optical flow algorithm.

Furthermore, for the LK algorithm to be applied efficiently on long sequences, our implementation uses an image discretization into cells – see below – and refills each cell whenever the number of tracked points inside it goes below a given threshold. This avoids to refill globally the image with new corners and lose temporal state variations everywhere in the image at this moment.

Estimating the distribution of velocity amplitudes. Based on this sparse optical flow, we can first estimate the variance parameter $\sigma_m^2(\mathbf{v}_t, \mathbf{r})$ of $p(v_{t+1}^m | \mathbf{v}_t, \mathbf{r}_t)$, by calculating the empirical variance of consecutive velocity magnitudes.

Estimating the distribution of velocity orientations. As for the second term of Eq. 1, we rely on a discrete representation of the (v^θ, \mathbf{r}) space. We discretize the image into $D \times D$ pixels-wide cells (in our experiments, $D = 30$) and the set of possible orientations into other 8 cells. Then, we build, for each of these 3D cells in the (v^θ, \mathbf{r}_t) space, a *discrete* representation of $p(v_{t+\tau}^\theta, \mathbf{r}_{t+\tau(v^m)} | \mathbf{v}_t, \mathbf{r}_t)$. As space is discretized into cells, it is difficult to consider in that case motion

information at time horizon 1, as most tracked points stay in the same cell; Instead, we collect temporal information, for each cell, at a time horizon that would span some significant motion towards a neighbour cell. A good choice for this time interval is $\tau(v^m) = \frac{D}{v^m}$, that represents the average time to cross the cell along a straight line. Hence, if a point is tracked by the LK algorithm, with velocity v^m at frame t , it contributes to the histogram at the 3D cell $(v_t^\theta, \mathbf{r}_t)$, into the bin (among $72 = 8 \times 9$ bins) corresponding to the observed $(v_{t+\tau(v^m)}^\theta, \mathbf{r}_{t+\tau(v^m)})$. This process is illustrated by Fig. 1, for a single point (in blue) tracked over a few frames. The central green dot gives a contribution for the cell representing its current state, based on its position $\tau(v^m)$ frames later. The corresponding bin is also depicted in green.

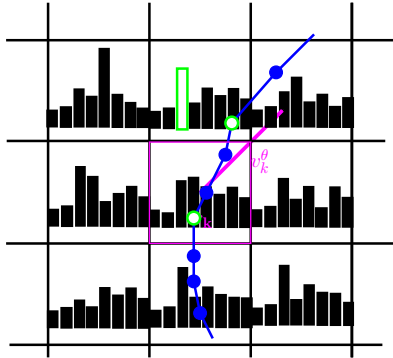


Fig. 1. Priors for the transition model by discretizing the image+orientations space. For one cell c_k , characterized by $(v_k^\theta, \mathbf{r}_k)$ (magenta), we model the distribution of $(v_{k'}^\theta, \mathbf{r}_{k'})$, i.e. the orientations and cells reached by the target after a “significant” amount of time $\tau(v^m)$. In green, one contribution from a point tracked on the blue path.

From now on, we will refer to each cell as $\mathbf{c}_k = (v_k^\theta, \mathbf{r}_k)$, with k indexing the set of cells, and $\mathbf{h}(\mathbf{c}_k)$ being the normalized histogram formed at that cell.

3.2 Regularization of Velocity Orientation Distributions

Let B be the number of bins in the histogram, $\mathbf{h}(\mathbf{c}_k)_i$ the value of the i -th bin in the normalized histogram, and $\log \mathbf{h}(\mathbf{c}_k)$ the histogram made of the logs of the entries of $\mathbf{h}(\mathbf{c}_k)$, i.e. $[\log \mathbf{h}(\mathbf{c}_k)]_i \stackrel{\text{def}}{=} \log [\mathbf{h}(\mathbf{c}_k)_i]$.

Now, as we will see later, given the high dimension of the space on which are defined all local distributions (72) and the relatively few data we have to estimate them, we need to estimate a regularized version $\mathbf{l}(\mathbf{c}_k)$ of $\log \mathbf{h}(\mathbf{c}_k)$ with the optimization scheme described hereafter.

The objective function we will minimize is the following one:

$$U(\mathbf{l}) = \frac{1}{2} \sum_{k \in K} \sum_{i,j} \mathbf{W}_{ij} ([\mathbf{l}_i(\mathbf{c}_k) - \log \mathbf{h}_j(\mathbf{c}_k)]^2 + \lambda \sum_{\mathbf{l} \in N(k)} [\mathbf{l}_i(\mathbf{c}_k) - \mathbf{l}_j(\mathbf{c}_l)]^2). \quad (2)$$

We can make several observations on this objective function. First, note that the searched optimum should reside in the manifold made of the concatenation of the logs of normalized histograms, i.e. for any cell \mathbf{c}_k , $\mathbf{l}(\mathbf{c}_k) \in \mathcal{L} = \{\mathbf{l} \in \mathbb{R}^B \text{ s.t. } \sum_{i=1}^B e^{l_i} = 1\}$. Then, the first term of the function is a *data* term, that fits the l 's to the empirical data. The second term is a smoothness constraint, that makes the histogram at one cell \mathbf{c}_k similar to the cells \mathbf{c}_l of its neighbourhood $N(k)$. These neighbourhoods include the spatial vicinity, of course, and the angular vicinity. Also note that λ is the regularization factor, and $K \subset \mathbb{N}$ is the set of considered cells. We will see below that we do not necessarily consider all the cells, but only the most informative ones.

Last, the terms \mathbf{W}_{ij} are weights that encode the similarity between different, but close histograms bins, i.e. they soften the binning effect; they also include the particular fact that angle histograms are cyclic. For this purpose, we use the von Mises distribution with the angle between the direction vectors, which represent the bins i and j .

By developing a bit more the expression of $U(\mathbf{l})$ in Eq. 2,

$$U(\mathbf{l}) = C + \frac{1}{2} \sum_{\mathbf{c} \in K} (\mathbf{l}(\mathbf{c}_k)^T \mathbf{W}^{(1)} \log \mathbf{h}(\mathbf{c}_k) + \sum_{\mathbf{l} \in N(\mathbf{k})} \mathbf{l}(\mathbf{c}_k)^T \mathbf{W}^{(2)}(\mathbf{c}_k, \mathbf{c}_l) \mathbf{l}(\mathbf{c}_l)),$$

with C a constant, and

$$\mathbf{W}_{ij}^{(1)} = -2\lambda \mathbf{W}_{ij},$$

$$\mathbf{W}_{ij}^{(2)}(\mathbf{c}_k, \mathbf{c}_l) = \begin{cases} \begin{cases} -2\lambda \mathbf{W}_{ij} & \text{if } k \neq l \\ (1 + 2V\lambda) \sum_l \mathbf{W}_{il} - 2\lambda \mathbf{W}_{ii} & \text{if } i = j \\ -2\lambda \mathbf{W}_{ij} & \text{otherwise.} \end{cases} & \text{if } k = l. \end{cases}$$

Expressed this way, the problem of Eq. 2 is quadratical in the vector formed by all \mathbf{l} and can be solved with classical numerical optimization techniques. We adopted a Gauss-Seidel scheme on the gradient of the objective function, in which after each iteration the new estimate for \mathbf{l} is projected on the manifold $\mathcal{L} \subset \mathbb{R}^{|K|B}$ (i.e. re-normalization) [9]. In our experiments, approximately 40 iterations were necessary to ensure convergence, depending on the value chosen for the regularization parameter λ .

An example of regularized prior is shown (partially) in Fig. 2. It sums up the histograms $\mathbf{l}(\mathbf{c}_k)$ obtained with the optical flow data collected among several of the PETS'2009 video sequences, which is a dataset of common use in the evaluation of tracking algorithms [10]. The discretization being 3D, we depicted, for one particular initial orientation $v_i^\theta = \frac{\pi}{2}$ (i.e., a downward motion, in white), the main orientation observed at each of the potential arrival cell. Each image corresponds to a potential arrival cell (which one is indicated by the bold arrow in the bottom-left corner), and the colored arrow at each cell indicates the most

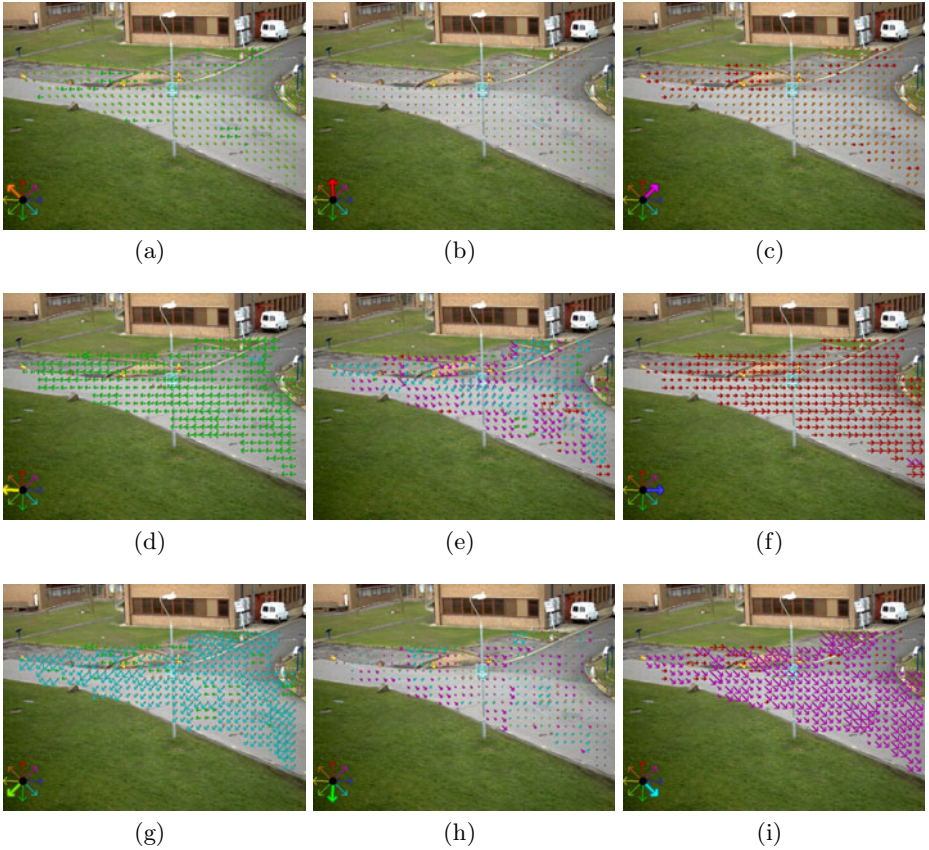


Fig. 2. A glimpse on the learned distribution $p(v_{t+1}^\theta, \mathbf{r}_{t+1} | \mathbf{v}_t, \mathbf{r}_t)$, for $v_t^\theta = \frac{\pi}{2}$ (downward motion, in white): each image corresponds to a spatial displacement (i.e., \mathbf{r}_{t+1}) indicated by the bold arrow on the bottom-left corner, and the colored arrows give the orientation (i.e., v_{t+1}^θ) with the highest bin among those corresponding to this \mathbf{r}_{t+1} .

frequent orientation for that cell, its length indicating the value of the corresponding bin. A remarkable element, justifying not to use too simple motion models, is that for this vertical orientation, there is nearly no support for the cell immediately downward (Fig. 2(h)). From this camera, vertical trajectories are seldom observed, downward velocities correspond generally to trajectories oriented to the south-west (Fig. 2(g)) or to the south-east (Fig. 2(i)). Finally, in Fig. 3, we show more synthetic views of the transition distributions, through the sole maximal bins attained, for a given initial orientation, which is the upward direction in this case. The three arrows displayed correspond to the initial orientation, the direction to the attained new cell, and the final orientation. On the left side, we depict the non-regularized version, and on the right side its regularized counterpart. As in Fig. 2, one can note that vertical motions are generally not followed along straight line: They tend to bend either to the left



Fig. 3. Raw and regularized transition fields for an upward initial direction ($v_t^\theta = -\frac{\pi}{2}$). We depict, before and after regularization, the 3D cell having the maximal bin value, with a first arrow pointing to \mathbf{r}_{t+1} and a second arrow having orientation v_{t+1}^θ .

(for example in the bottom-right corner) or to the right (upper-right corner). This reflects the physical constraints present in the scene for pedestrians.

To complete the regularization process, we have also tried to fill in incomplete information. As written before, because of the lack of data, some cells have no information at all, or nearly no information: For example, areas that are not accessible to pedestrians or areas hidden behind occluding clutter, such as the central pole at Fig. 3. Hence, we adopted a strategy to fill in these gaps: First we generate a binary image from data with not enough information and we decompose it into convex regions. We estimate the boundaries of two regions that are spatially close and sample a point in each region. The pair of points must have an Euclidean distance inferior to a threshold (i.e. we assume that points far away are not physically attainable). Then we sample an orientation at each point and interpolate positions/orientations by a cubic polynomial. Finally, we generate histogram entries according to the polynomial curvature. This strategy helps completing the missing information in some areas. We will see in Section 5, that it helps preserving the trajectories continuity.

4 Using the Learned Motion Models during Tracking

With this acquired knowledge about how pedestrians tend to move in the scene, we can now define a corresponding probabilistic motion model for performing target tracking, i.e. not only the low-level tracking of a single point as LK does, but the one of a complex target. One of the most flexible techniques for performing such a tracking task is the particle filter (PF), as it allows to integrate in an elegant way a probabilistic model on what we should observe from the target in the current image (observation model), and a probabilistic model on what we know about how this targets moves (motion model). Above all, it makes no strong assumption about these probability distributions, the only constraints

being, in the simplest form of PF, that one could evaluate the observation model at any state, and that one could sample the motion model from any state.

4.1 Particle Filter-Based Visual Tracking

Formally, we will index the targets to track with indices m , and associate one individual filter to each target. Any of these filters estimates the $4 - D$ Markov state of its associated target at time t , $\mathbf{X}_t^m = (\mathbf{r}_t^m, \mathbf{v}_t^m)^T$ from the sequence of images $\mathbf{I}_1, \dots, \mathbf{I}_t$. In practice, for the problem of tracking, the state contains the target Region of Interest position and its velocity. We will suppose, as other authors [11], that we have a rough knowledge of the position of the camera with respect to the scene, so that, at one possible position for the target in the image, the corresponding scale of the target bounding box is known. To do the target state estimation, the PF uses the recursive application of Bayes rule, that leads to

$$p(\mathbf{X}_t^m | \mathbf{I}_1, \dots, \mathbf{I}_t) = p(\mathbf{I}_t | \mathbf{X}_t^m) \int_{\mathbf{X}_{t-1}^m} p(\mathbf{X}_t^m | \mathbf{X}_{t-1}^m) p(\mathbf{X}_{t-1}^m | \mathbf{I}_1, \dots, \mathbf{I}_{t-1}) d\mathbf{X}_{t-1}^m.$$

To get an approximate representation of this posterior from the previous equation, PF uses a Monte-Carlo approach, with a set of weighted samples (or “particles”) from the posterior distribution, $\{(\mathbf{X}_t^{m,(n)}, \omega_t^{m,(n)})\}_n$, where n indexes the particles [12]. At each time t , these samples are generated from a proposal distribution $q(\mathbf{X}_t^m | \mathbf{X}_{t-1}^m, \mathbf{I}_t)$, and their weights are recursively updated so that they reflect how much the particles consecutive states evaluate under the posterior distribution. When, as we do here, the proposal distribution is precisely the probabilistic motion model, i.e.

$$q(\mathbf{X}_t^m | \mathbf{X}_{t-1}^m, \mathbf{I}_t) = p(\mathbf{X}_t^m | \mathbf{X}_{t-1}^m),$$

then the weight update rule is simply $\omega_t^{m,(n)} = p(\mathbf{I}_t | \mathbf{X}_t^{m,(n)}) \omega_{t-1}^{m,(n)}$. A resampling step is applied whenever the number of significative particles (measured by $\frac{1}{\sum_n (\omega_t^{m,(n)})^2}$) becomes inferior to a threshold.

Here, we have used such a particle filter with a rather simple and common observation model, based on existing works on visual tracking [13]. It combines two main visible features of the target, its color distribution and its motion distribution along the video sequence. The first feature (color) is encoded through 3 H, S, V histograms defined in each of two sub-regions defined over the target region of interest, the upper one and the lower one, for a total of six histograms. The choice of dividing the target region into two comes with the idea of associating some spatial information to the appearance model, in addition of the pure color data. As pedestrians clothes have generally quite different color distributions in their upper and lower parts, this observation model gives a much more discriminative power. The second feature (motion) is also encoded into a histogram, that contains the distribution of grey value differences between the current and the previous images in the sequence. For all of these histograms (color and motion),

we store a reference histogram at the first frame where the target is detected, which is further updated along the video sequence with a simple exponential decay rule. As all of these features have the same form, the overall likelihood takes the form

$$p(\mathbf{I}_t | \mathbf{X}_t^m) \propto \prod_f \exp \left(-\frac{d^2(\mathbf{h}_f^m(\mathbf{X}_t^m), \mathbf{h}_f^{m*})}{2\sigma_f^2} \right),$$

where d is a distance between histograms (Bhattacharya distance), \mathbf{h}_f^{m*} is the reference histogram for feature f (among 7 features) and $\mathbf{h}_f^m(\mathbf{X}_t^m)$ is the histogram computed at the target position \mathbf{X}_t^m . σ_f^2 the variance on the error on the histogram distance, which is a set of parameters for the algorithm.

4.2 Using the Learned Motion Model

Now, given the learned motion model, we define a proposal definition based on it. We will refer to the underlying state transition model as $\pi(\mathbf{X}_{t+1}^m | \mathbf{X}_t^m)$. To draw samples from this distribution, conditionally to the state at t of the particle n , $\mathbf{X}_t^{m,(n)}$, we use the following steps:

1. Determine the 3D cell $\mathbf{c}_k = (v_k^\theta, \mathbf{r}_k)$ (position+velocity direction) the particle $\mathbf{X}_t^{m,(n)}$ corresponds to;
2. From the learnt distribution, $p(v_{k'}^\theta, \mathbf{r}_{k'} | v_k^\theta, \mathbf{r}_k)$, sample a cell $\mathbf{c}_{k'}$ where the tracked target could probably go after some time, with its future orientation;
3. Sample a velocity amplitude $v_{t+1}^{m,(n)}$ from $p(v_{t+1}^m, |\mathbf{v}_t^{m,(n)}, \mathbf{r}_t^{m,(n)})$;
4. From the pairs of cells, and their corresponding orientations, build a cubic curve joining the center cells and tangent to the initial and final orientations (i.e. Hermite interpolation);
5. Translate this curve on the 2D position of the particle $\mathbf{X}_t^{m,(n)}$ and sample a point on it around the position given by the velocity amplitude $v_{t+1}^{m,(n)}$. To make any neighbor configuration reachable from one initial configuration, we also add a lateral noise along the polynomial curve.

This way, we can sample from $\pi(\mathbf{X}_{t+1}^m | \mathbf{X}_t^m)$. Now, we may also define a more classical constant velocity model, that we will refer to as $p(\mathbf{X}_{t+1}^m | \mathbf{X}_t^m)$, and a mixture regulated by a parameter γ .

$$\pi'(\mathbf{X}_{t+1}^m | \mathbf{X}_t^m) = \gamma p(\mathbf{X}_{t+1}^m | \mathbf{X}_t^m) + (1 - \gamma) \pi(\mathbf{X}_{t+1}^m | \mathbf{X}_t^m). \quad (3)$$

Through this mixture, we have evaluated several ways to use the prior: as a fixed mixture proposal ($\gamma = \frac{1}{2}$), or as a proposal to be used whenever the filter undergoes difficulties, e.g. because of occlusions. In that case, the prediction from the constant velocity model is made risky, since the state estimation is poor. In this last case, the coefficient γ weighting the two distributions is a quality measure evaluating the current estimation by the particle filter. One simple way to define it is as the average likelihood after the observation model is taken into account: $\sum_n p(\mathbf{I}_t | \mathbf{X}_t^{m,(n)}) \omega_t^{m,(n)}$.

Table 1. Performance evaluation of tracking. Results using PETS’2009 and CAVIAR datasets: First row (of each table) show the results for a classic SIR particle filter, with a constant velocity motion model. The last three rows are our results using motion prior with raw data, regularized data and with the strategy of filling incomplete data. Note : All results are the median value of 30 experiments. On the right, γ is taken variable, as the quality measure of the filter; on the left, it is fixed.

PETS’2009, $\gamma = \frac{1}{2}$				PETS’2009, γ as quality			
SFDA	ATA	N-MODP	MOTP	SFDA	ATA	N-MODP	MOTP
0.40	0.42	0.51	0.51	0.40	0.42	0.51	0.51
0.36	0.39	0.49	0.49	0.38	0.43	0.50	0.50
0.39	0.45	0.50	0.51	0.41	0.47	0.52	0.52
0.40	0.46	0.52	0.52	0.43	0.48	0.54	0.53

CAVIAR, γ as quality			
SFDA	ATA	N-MODP	MOTP
0.10	0.10	0.23	0.58
0.14	0.12	0.37	0.62
0.14	0.13	0.40	0.68
0.15	0.13	0.40	0.70

5 Experimental Results

We evaluated our proposal on two public datasets: CAVIAR and PETS’2009. The first one has a ground truth but the second one has not, so we have manually generated one for the occasion. To evaluate our results quantitatively, we used a now standard methodology developed by the PETS community [10]. Tracking quality is not always easy to quantify, hence several metrics have been proposed, and we use four of them: (1) Normalized Multiple Object Detection Precision (N-MODP), which reflects the target detection rate and precision; (2) Multiple Object Tracking Precision (MOTP), that measures the tracks precision; (3) Sequence Frame Detection Accuracy (SFDA), and (4) Average Tracking Accuracy (ATA), which measures tracks precision but takes more into account the shortening of trajectories. These four indicators take values in the interval $[0, 1]$ (1 being for high quality).

The table 1 presents some results obtained for these indicators, on the two aforementioned datasets. In each case, the first row gives the indicator levels obtained with a classical SIR particle filter using a constant velocity motion mode. Then, the next three rows give results for the same SIR trackers using (1) the raw motion fields as probabilistic motion models; (2) the regularized motion fields and (3) the regularized motion fields with the hole filling strategy mentioned above. In all cases the results are improved, in particular for the first two indicators, that are sensible to the continuity of trajectories. Note that the results for the CAVIAR sequence are low, and this can be explained by the low sensibility of motion detection in that case (we used the OpenCV motion detector), that makes the tracker initialization long to occur, and makes the indicator levels drop in that case. Also note that we have compared two policies for the choice of γ , the policy setting it as a fixed value ($\gamma = \frac{1}{2}$, left column), and the policy of using the filter average likelihood as a measure. The best results are

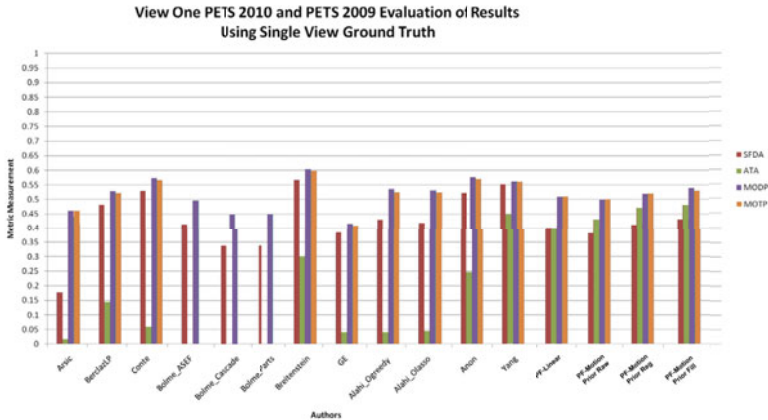


Fig. 4. Performance evaluation of tracking proposed by other authors and our proposal in set S2.L1 (view 1) of PETS 2009 dataset, for the four quality indicators. The last four results use our tracking system with linear motion model, motion prior model with raw data, a motion prior model with regularized data and a motion prior model with the strategy of filling zones with incomplete information. Other authors results have been reported in [10].

observed when taking the quality measure, i.e. when this proposal is really used when the filter is not tracking well the target. Last, Fig. 4 shows a comparison of our own results for the PETS’2009 sequence with results of other authors from PETS 2009 and 2010 conferences. As it can be noticed, the overall performance of our approach for the four indicators locates it among the best entries.

6 Conclusions and Future Work

We have presented a particle-based approach for visual tracking in video-surveillance sequences that relies, more than on a particularly efficient observation model, on a probabilistic motion model that is learned from sequences grabbed by the same camera. This way, the particle filter sampling is done in areas corresponding to paths that are much more likely to be taken by pedestrians than what would be obtained with more traditional motion models (for example, the constant velocity motion model). Experiments on classical datasets of video-surveillance data and evaluation through standard indicators have shown that such an approach could be quite promising in obtaining better tracking results (where the term “better” may cover a complex reality, which is the reason why different tracking quality indicators have been used).

We plan several extensions for enhancing particle filter-based tracking: first, we plan to incorporate more scene-related elements in the regularization framework, i.e. to integrate boundaries (i.e. road limits) that should be taken into account, so as not to smooth motion fields through these boundaries; second, we

plan to use this framework in an incremental way, so that no explicit learning phase would be required, and the model could be updated on a regular basis.

References

1. Leibe, B., Schindler, K., Van Gool, L.: Coupled detection and trajectory estimation for multi-object tracking. In: Proc. Int. Conf. on Computer Vision (2007)
2. Kuo, C.H., Huang, C., Nevatia, R.: Multi-target tracking by on-line learned discriminative appearance models. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2010), pp. 685–692 (2010)
3. Fragkiadaki, K., Shi, J.: Detection free tracking: Exploiting motion and topology for tracking and segmenting under entanglement. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, CVPR 2011 (2011)
4. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2005), pp. 878–885 (2005)
5. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR 2000), vol. 2, pp. 142–149 (2000)
6. Sidenbladh, H., Black, M.J., Sigal, L.: Implicit Probabilistic Models of Human Motion for Synthesis and Tracking. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part I. LNCS, vol. 2350, pp. 784–800. Springer, Heidelberg (2002)
7. Bouguet, J.Y.: Pyramidal implementation of the Lucas Kanade feature tracker description of the algorithm. In: USENIX Technical Conference (1999)
8. Shi, J., Tomasi, C.: Good features to track. In: Int. Conf. on Computer Vision and Pattern Recognition (CVPR 1994), pp. 593–600 (1994)
9. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, New York (2006)
10. Ellis, A., Ferryman, J.: Pets2010 and pets2009 evaluation of results using individual ground truth single views. In: Proc. of the IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS), pp. 135–142 (2010)
11. Madrigal, F., Hayet, J.: Multiple view, multiple target tracking with principal axis-based data association. In: Proc. of the IEEE Int. Conf. on Advanced Video and Signal based Surveillance, AVSS (2011)
12. Doucet, A., De Freitas, N., Gordon, N. (eds.): Sequential Monte Carlo methods in practice. Springer, Heidelberg (2001)
13. Perez, P., Vermaak, J., Blake, A.: Data fusion for visual tracking with particles. Proc. of the IEEE 92(3), 495–513 (2004)

CT-MR Image Registration in 3D K-Space Based on Fourier Moment Matching

Hong-Ren Su¹ and Shang-Hong Lai^{1,2}

¹Institute of Information Systems and Applications

²Department of Computer Science

National Tsing Hua University, Hsinchu, Taiwan

d9765805@oz.nthu.edu.tw, lai@cs.nthu.edu.tw

Abstract. CT-MRI registration is a common processing procedure for clinical diagnosis and therapy. We propose a novel K-space affine image registration algorithm via Fourier moment matching. The proposed algorithm is based on estimating the affine matrix from the moment relationship between the corresponding Fourier spectrums. This estimation strategy is very robust because the energy of the Fourier spectrum is mostly concentrated in the low-frequency band, thus the moments of the Fourier spectrum are robust against noises and outliers. Our experiments on the real CT and MRI datasets show that the proposed Fourier-based registration algorithm provides higher registration accuracy than the existing mutual information registration technique.

Keywords: Multi-modal image registration, Fourier moments, CT, MRI.

1 Introduction

Multi-modal medical image registration has been an important research topic in medical imaging due to its great value in clinical applications [1]. One of the most important multi-modal image registrations is the CT-MRI registration [2], which is widely used in minimally invasive surgery and 3D conformal and intensity-modulated radiotherapy (IMRT).

Computed tomography (CT) [3] is commonly used in routine diagnosis since it provides higher spatial accuracy and hard-tissue contrast, such as the bone tissue. It also has the electron density information necessary for radiotherapy (RT) but its disadvantage is the poor soft-tissue contrast, which causes diagnostic problems in soft tissue, like the tumor or nerve. In contrast, magnetic resonance image (MRI) [3] provides superior soft-tissue contrast and visualization in axial, sagittal and coronal planes, allowing better 3D representation for diagnosis. However MRI is not sensitive to the hard tissue and suffers from artifacts at interfaces between bone and air. A feasible and useful strategy [4] for clinical application is to combine the complementary advantages of CT and MRI. Thus, an accurate image registration between CT and MR images is essential for diagnosis and therapy.

Existing methods used in hospitals for CT-MR registration are mark-and-link [2] or mutual information (MI) based image registration techniques [5]. "Mark-and-link" registration needs matching points between CT and MRI selected by specialists. It spends a lot of money and time but the accuracy is not robust and matching points is not easy selected in a large three dimension data. An automatic and efficient method is the MI technique, which is suitable for multi-modal registration problems and has become a standard algorithm [6, 7, 8]. It is based on minimizing the mutual information cost function between the corresponding voxel image intensities and useful because two modal images have similar intensity distribution. However, MI has an accuracy limitation [4, 8] and cannot always deal with the data well so that many modified methods based on MI have been proposed to improve the accuracy [7].

In this paper, we focus on the CT-MRI registration problem, which is the structure-to-structure image registration. Structure medical images [3], such as CT and MRI, have clear edge information allowing distinguishing one organ area from the others. Thus, there is similar edge information between CT and MRI for human which can be used in the registration. We propose a novel and robust k-space image registration algorithm by using edge information for CT-MRI registration. The K-space is a concept of Fourier space that is well known in medical imaging [9]. The relation between K-space data and image data is the Fourier Transformation.

The K-space registration [10, 11] is another intensity-based registration technique that is robust to noise with low computation complexity. It is more accurate and reliable to use the multi-layer fractional Fourier transform [12] than the traditional Fourier transform. However, all the previous image registration methods in K-space can only deal with translation, rotation and scaling, i.e. scaled rigid transformation. In addition, it is not straightforward to extend the K-space registration approach to higher-dimensional image registration. Therefore, there has not been much effort of applying the Fourier approach for medical image registration, especially for CT-MRI registration, which involves three-dimensional image registration.

In this paper, we propose a novel K-space affine image registration algorithm for CT-MRI registration. The proposed algorithm is based on the fact that the affine transform between two images corresponds to a related affine transform between their Fourier spectrums [13], whose energies are normally concentrated around the origin in the frequency domain. Thus, the moments for the corresponding Fourier spectrum distributions can be calculated as probability density function. In short, the proposed affine registration algorithm is based on minimizing the affine relationship between the moments for the Fourier spectrums of the two images. In addition, we further extend the algorithm to solve CT and MRI registration problem by representing the 3D image data as point sets in the 3D space as a binary image. The image registration problem is converted into a point-set registration problem in a 3D space. We introduce an appropriate distance weighting scheme determined from the shortest distance between the affine transformed point sets in the binary image to achieve better robustness.

To the best of our knowledge, this is the first work that solves the affine image registration problem by using moment matching in the K-space, and the proposed algorithm can also be applied to higher dimension.

The rest of this paper is organized as follows. In the next section, we briefly review the affine transform relationship between two images in the spatial domain has corresponding affine relationship in the Fourier domain. In section 3, we propose a novel Fourier-based algorithm for robust affine image registration. Experimental results on fMRI motion correction by using the proposed algorithm and some previous methods are given in section 4. Finally, we conclude this paper in the last section.

2 Fourier-Based Affine Image Registration

There have been several Fourier-based methods proposed for image registration in the past [10, 11, 12, 14, 15]. The Fourier-based methods have the advantages that it is efficient and can handle large motion. Recently, some modified Fourier-based image registration methods, like log-polar method [14], multi-layer Fourier transform (MLFFT) [12] and the phase correlation method [15], have been proposed to improve the registration precision and the alignment range. However, all the Fourier-based image registration methods can only deal with rigid transformation and cannot be easily extended to higher dimensions. In this paper, we propose the first Fourier-based affine registration algorithm by estimating the affine transformation from the corresponding Fourier spectrums of the two images.

2.1 Two-Dimensional Case

Consider two images $g(x,y)$ and $h(x,y)$ and they are related by an affine transformation [13], i.e. $h(x,y) = g(ax+by+c,dx+ey+f)$. Assume the Fourier transforms of the image functions $g(x,y)$ and $h(x,y)$ be denoted by $G(u,v)$ and $H(u,v)$, respectively. Then, we can derive the following affine relationship between their Fourier transforms $G(u,v)$ and $H(u,v)$ given as follows []:

$$G(u,v) = \frac{1}{|\Delta|} e^{i\frac{[(ec-bf)u+(af-cd)v]}{\Delta}} H\left(\frac{eu-dv}{\Delta}, \frac{-bu+av}{\Delta}\right) \quad (1)$$

By letting $u'=(eu-dv)/\Delta$ and $v'=(-bu+av)/\Delta$, we have the relationship $u=au'+dv'$ and $v=bu'+ev'$. Taking the absolute values on both sides of eq. (1), we have the affine transformation relationship between the spectrums $|G(u,v)|$ and $|H(u,v)|$ as follows:

$$|G(u,v)| = \frac{1}{|\Delta|} |H(u',v')| \quad (2)$$

where

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} a & d \\ b & e \end{bmatrix} \begin{bmatrix} u' \\ v' \end{bmatrix}, \Delta = \begin{vmatrix} a & b \\ d & e \end{vmatrix} = ae - bd \quad (3)$$

From eq. (2) and (3), we can see the Fourier spectrums of two images related by an affine transform are also related by the corresponding affine transform. If we can estimate the four affine parameters in eq. (3) from the two Fourier spectrums, the remaining two affine parameters can be computed by transforming image $g(x,y)$ with the transformation $g'(x,y)=g(ax+by,dx+ey)$ and determining the translation vector (c,f) between $g'(x,y)$ and $h(x,y)$ from their cross power spectrum. To be more specific, the translation vector is determined as follows:

$$(c, f) = \arg \max_{(x,y)} \text{real}(IFT \left\{ \frac{G'(u,v)H^*(u,v)}{|G'(u,v)H^*(u,v)|} \right\}) \tag{4}$$

where $G'(u,v)$ is the Fourier transform of $g'(x,y)$, $H^*(u,v)$ is the complex conjugate of $H(u,v)$, and IFT denotes the inverse Fourier transform operator.

2.2 Three-Dimensional Case

In the above section, the 2D model of the affine transform relationship between image and Fourier domain is described. It is easy to extend it to 3D or higher dimension model. Consider two 3D image functions $g(x,y,z)$ and $h(x,y,z)$, which are related by an affine transformation, i.e. $h(x,y,z) = g(ax+by+cz+d,ex+fy+gz+h,ix+jy+kz+l)$. Assume the Fourier transforms of $g(x,y,z)$ and $h(x,y,z)$ be denoted by $G(u,v,w)$ and $H(u,v,w)$, respectively. Then, we can derive the following affine relationship between the Fourier transforms $G(u,v,w)$ and $H(u,v,w)$ given as follows:

$$|G(u, v, w)| = \frac{1}{|\Delta|} |H(u', v', w')| \tag{5}$$

where

$$\begin{bmatrix} u \\ v \\ w \end{bmatrix} = \begin{bmatrix} a & e & i \\ b & f & j \\ c & g & k \end{bmatrix} \begin{bmatrix} u' \\ v' \\ w' \end{bmatrix}, \Delta = \begin{vmatrix} a & b & c \\ e & f & g \\ i & j & k \end{vmatrix} \tag{6}$$

Similar to eq. (4) in the 2D case, the translation vector (d,h,l) between $g(x,y,z)$ and $h(x,y,z)$ can be determined from their cross power spectrums.

From the above discussion on 2D and 3D affine transforms, we can see the main problem to estimate the affine registration parameters based on the Fourier domain approach is how to estimate the affine parameters (a, b, d, e) in the 2D case and $(a, b, c, e, f, g, i, j, k)$ for the 3D case from a pair of Fourier spectrums as given in eq. (2). In this paper, we propose to estimate these affine parameters via the moment matching technique proposed by Ho et al. [16], which will be described subsequently in the next section.

2.3 Moment Matching Approach to Estimating Affine matrix

Let the Fourier spectrums of the 2D image functions $f_1(x,y)$ and $f_2(x,y)$ be denoted by $F_1(u,v)$ and $F_2(u,v)$, respectively. If $f_1(x,y)$ and $f_2(x,y)$ are related by an affine transformation, then their Fourier spectrums are also related by the corresponding affine transform, i.e. $|F_1(u, v)| = |F_2(u', v')|/\Delta$ with the relation between (u, v)

and (u', v') given in eq. (3). To determine the affine parameters, we employ the moment matching technique to the Fourier spectrums $|F_1(u, v)|$ and $|F_2(u, v)|$.

The $(\alpha+\beta)$ -th moment for the Fourier spectrum $|F_k(u, v)|$, $k=1$ or 2 , is defined as

$$m_{\alpha,\beta}^k = \iint u^\alpha v^\beta |F_k(u, v)| dudv \quad (7)$$

By coordinate substitution, we can derive the following equation

$$m_{\alpha,\beta}^2 = \iint (au' + dv')^\alpha (bu' + ev')^\beta |F_2(u', v')| du' dv' \quad (8)$$

Thus, we have the following relationship for the first-order moments.

$$\begin{bmatrix} m_{1,0}^1 \\ m_{0,1}^1 \end{bmatrix} = \begin{bmatrix} a & d \\ b & e \end{bmatrix} \begin{bmatrix} m_{1,0}^2 \\ m_{0,1}^2 \end{bmatrix} \quad (9)$$

For the second-order moments, we can derive the following relationship.

$$\begin{bmatrix} m_{2,0}^1 \\ m_{1,1}^1 \\ m_{0,2}^1 \end{bmatrix} = \begin{bmatrix} a^2 & 2ad & d^2 \\ ab & ae + bd & de \\ b^2 & 2be & e^2 \end{bmatrix} \begin{bmatrix} m_{2,0}^2 \\ m_{1,1}^2 \\ m_{0,2}^2 \end{bmatrix} \quad (10)$$

The 2D affine parameters (a, b, d, e) can be estimated by minimizing the total errors associated with the constraints in eq. (9) and (10) in a least-squares estimation framework [16].

The 3D affine parameters $(a, b, c, e, f, g, i, j, k)$ can be estimated by similar moment matching technique from the moments of the Fourier spectrums $F_1(u, v, w)$ and $F_2(u, v, w)$. The $(\alpha+\beta+\gamma)$ -th moment for the Fourier spectrum $|F_n(u, v, w)|$ is defined as

$$m_{\alpha,\beta,\gamma}^k = \iiint u^\alpha v^\beta w^\gamma |F_k(u, v, w)| dudvdw \quad (11)$$

Thus, we have the following relationship for the first-order moments in the Fourier spectrums.

$$\begin{bmatrix} m_{1,0,0}^1 \\ m_{0,1,0}^1 \\ m_{0,0,1}^1 \end{bmatrix} = \begin{bmatrix} a & e & i \\ b & f & j \\ c & g & k \end{bmatrix} \begin{bmatrix} m_{1,0,0}^2 \\ m_{0,1,0}^2 \\ m_{0,0,1}^2 \end{bmatrix} \quad (12)$$

For the second-order Fourier moments, we can derive the following relationship.

$$\begin{bmatrix} m_{2,0,0}^1 \\ m_{0,2,0}^1 \\ m_{0,0,2}^1 \\ m_{1,1,0}^1 \\ m_{1,0,1}^1 \\ m_{0,1,1}^1 \end{bmatrix} = \begin{bmatrix} a^2 & e^2 & i^2 & 2ae & 2ai & 2ei \\ b^2 & f^2 & j^2 & 2bf & 2bj & 2fj \\ c^2 & g^2 & k^2 & 2cg & 2ck & 2gk \\ ab & ef & ij & af+be & aj+bi & ej+fi \\ ac & eg & ik & ag+ce & ak+ci & ek+gi \\ bc & fg & jk & bg+cf & bg+cf & fk+gj \end{bmatrix} \begin{bmatrix} m_{2,0,0}^2 \\ m_{0,2,0}^2 \\ m_{0,0,2}^2 \\ m_{1,1,0}^2 \\ m_{1,0,1}^2 \\ m_{0,1,1}^2 \end{bmatrix} \quad (13)$$

The relationship of the first-order and second-order Fourier moments, given in eq. (12) and (13), can be used for the least-squares estimation of the above nine 3D affine parameters.

3 Iterative Refinement Process by a Distance Weighting

In the previous section, we introduce a novel Fourier-based image registration by applying the moment matching technique to the Fourier spectrums of the image functions. To further improve the accuracy of the novel Fourier-based image registration algorithm, especially in the occlusive problem, we propose an iterative refinement process by introducing a distance weighting scheme into images, which is detailed in the following.

3.1 Canny Edge Image Data

Let a point set $p(\mathbf{x}) \in E$, which E is the set of the canny edge, $\mathbf{x} \in \mathbb{R}^n$, is extracted from an image $h(\mathbf{x})$. Then the image $h(\mathbf{x})$ is transferred into a binary image $B(\mathbf{x})$ with values of the pixels corresponding to the points p set to 1, and the rest set to 0.

$$B(\mathbf{x}) = \begin{cases} 1 & (\mathbf{x}) \in p \\ 0 & (\mathbf{x}) \notin p \end{cases} \quad (14)$$

3.2 Distance Weighting

The idea to improve the accuracy of the proposed Fourier-based affine image registration is to assign an appropriate weight to each point p in the binary image such that the points without correspondences have very small weights and the points with proper correspondences have high weights in the binary image. In the previous definition of the binary image B given in eq. (14), the function has binary values to indicate presence of data points. Since the point sets may contain some noise variation, we introduce a distance weighting to reduce the influence of some points without proper correspondences.

The distance for one data point $\mathbf{p} \in E_1$ in the binary image $B_1(\mathbf{x})$ to the other point set E_2 for the binary image $B_2(\mathbf{x})$ is defined as

$$d(\mathbf{p}, E_2) = \min_{\mathbf{q} \in E_2} \|\mathbf{p} - \mathbf{q}\| \quad (15)$$

Note that the distance for all data points in E_1 to E_2 can be efficiently computed by using the distance transform. Then, we define the weighting for each data point in E_1 to E_2 as follows:

$$w(\mathbf{p}, E_2) = \frac{\sigma^2}{\sigma^2 + d^2(\mathbf{p}, E_2)} \quad (16)$$

Thus, we can compute the weighting function Bw_1 for a binary image B_1 as follows:

$$Bw_1(\mathbf{x}) = \begin{cases} w((\mathbf{x}), E_2) & (\mathbf{x}) \in E_1 \\ 0 & (\mathbf{x}) \notin E_1 \end{cases} \quad (17)$$

Similarly, we can compute the weighting function Bw_2 for a binary image B_2 as follows:

$$Bw_2(\mathbf{x}) = \begin{cases} w((\mathbf{x}), E_1) & (\mathbf{x}) \in E_2 \\ 0 & (\mathbf{x}) \notin E_2 \end{cases} \quad (18)$$

In our algorithm, we first apply the Fourier-based affine image registration to the binary image B_1 and B_2 without using the distance weighting. The estimated affine transform is applied to all points in E_1 to update the point set E_1 . Then, the distance weighting is computed to produce the weighting functions Bw_1 and Bw_2 as eq. (17) and (18). The Fourier-based affine image registration algorithm is applied to find the affine transformation between Bw_1 and Bw_2 , and the affine estimation result is used to refine the affine registration. This refinement process is repeated several times until convergence.

The iterative registration is very crucial to the robustness of the registration. If the first step cannot provide robust registration, then the iterative refinement may not converge to the correct registration results. The proposed Fourier-based affine registration algorithm is robust even without the iterative refinement. It is because the Fourier transform of the image bring most of the energy to the low-frequency region in the Fourier domain, thus making the affine registration determined from the moments of the Fourier coefficients robust against noises and outliers.

3.3 Proposed Affine Image Registration Algorithm with Iterative Refinement

In this section, we summarize the proposed affine image registration algorithm. The detailed procedure is given as follows:

1. Generate the Canny edge as binary images B_1 and B_2 from the two images $h(\mathbf{x})$ and $g(\mathbf{x})$ as eq. (14).
2. Compute the discrete Fourier transforms of B_1 and B_2 via FFT.
3. Compute the first-order and second-order moments for the amplitude in Fourier spectrums of B_1 and B_2 from eq. (11)
4. Determine the affine parameters in matrix A by minimizing the least-square errors associated with the moment matching constraints given in eq. (12) and (13).
5. Transform B_1 with the affine transform with the estimated matrix A and the transformed data is denoted by B_1' .
6. Determine the translation vector t between B_1' and B_2 via the cross power spectrum method given in eq. (4).
7. Shift the map B_1' with the translation vector t and compute the distance weighting from eq. (17) to form the weighting function Bw_1 .
8. Repeat step 2 to step 7 with the binary image replaced by the weighting functions computed in the previous step until the changes in the affine transformation parameters are within a small threshold.

Note that the above affine image registration algorithm can be used in any high-dimensional image data.

4 Experimental Results

In our experiments, the CT and MRI 3D datasets in Vanderbilt Database [17], from the Retrospective Image Registration Evaluation (RIRE) project, are used for comparing the registration accuracy of different methods in CT to MRI registration. In our experiment, we compare the proposed Fourier-moment based image registration with the image registration program used in SPM8 [18] which is based on mutual information registration [19]. In the Vanderbilt Database, each patient has four 3D data sets, such as CT (512x512x28), MR-T1 (256x256x25), MR-T2 (256x256x25) and MR-PD (256x256x25). The data sets also provide manually annotated eight corresponding points for CT to MR-T1, CT to MR-T2 and CT to MR-PD. We calculate the average distances of the eight corresponding points in the image data after the CT-MR registration as the performance measure.

Table 1. Average distances of eight corresponding points between original data to the data after CR-MR registration

CT-MR registration	Proposed method	SPM8[18, 19]
CT to MR-PD	0.80 mm	0.89 mm
CT to MR-T1	0.76 mm	0.83 mm
CT to MR-T2	0.93 mm	1.07 mm

The proposed method is used for 3D CT-MR registration. Fig. 1 and 2 depict some examples of the images, and the corresponding edge maps and K-space representations. Fig. 1 shows the multi-modal images with their edge maps by Canny edge detection and the corresponding K-spaces. Table 1 shows the proposed method has smaller distances between the corresponding landmark points after registration than the mutual information based algorithm [19] used in SPM8 [18].

Fig. 2 shows the edge maps in one axial plane after CT-MR registration by using the proposed algorithm. CT image has higher resolution than MR image so that the CT-MR registration problem is not only a 3D multi-modal registration but also a 3D multi-resolution registration. We can see from Fig. 2 that the experiments for CT to MR-PD and CT to MR-T1 have better registration results.

For simulated affine image registration experiment, we used the brain atlas MRI data set from [20] and generate synthesized images with 50 random affine transformations from the T1 and T2 brain images. Some sample images used in this experiment are shown in Fig. 3. The average relative errors in affine registration obtained by using the proposed algorithm and the affine registration tool in the SPM software [18], which is based on the algorithm proposed in [19], are evaluated and shown in Table 2.

Note that the relative error for the affine matrix estimation is defined as

$$\text{Relative error} = \|\hat{A} - A\|_F / \|A\|_F \quad (19)$$

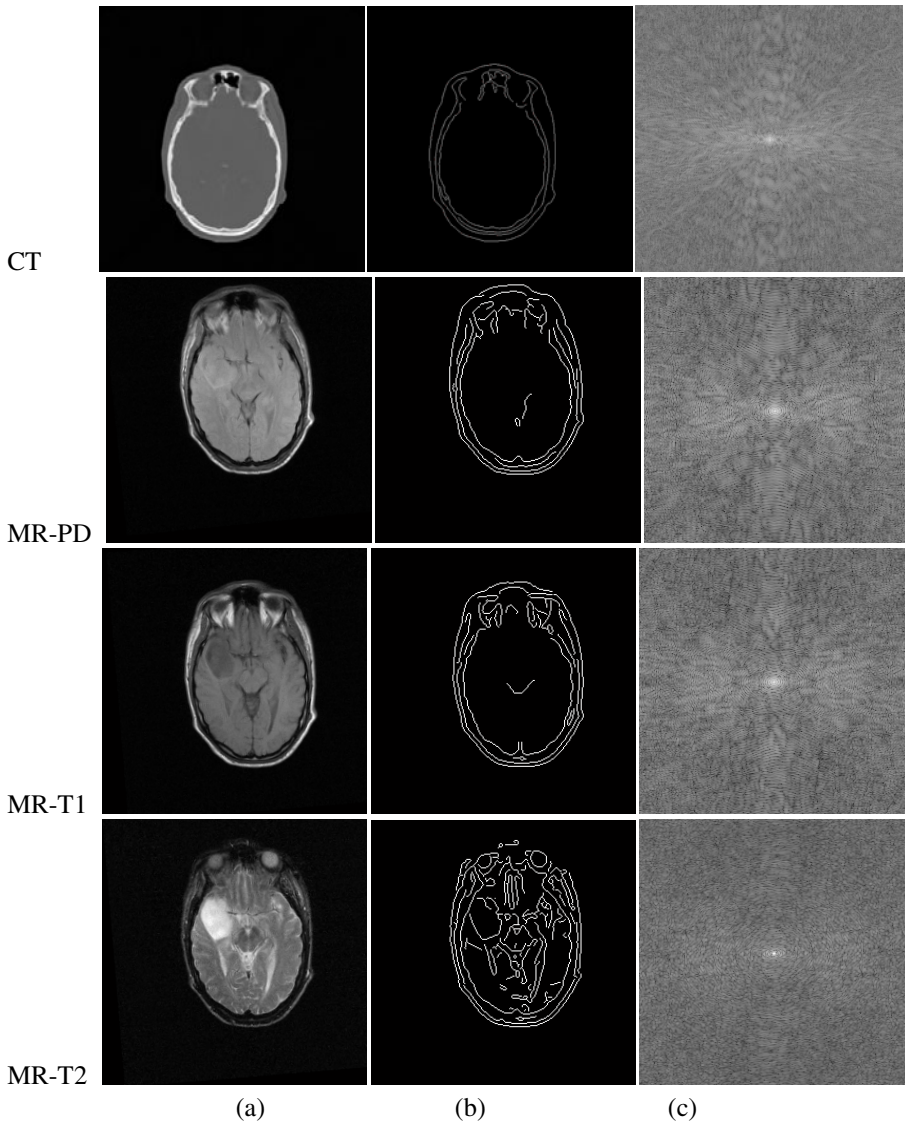


Fig. 1. Example images used in the CT-MR registration experiment from Vanderbilt Database: (a) original brain CT and MRI images, (b) canny edge maps from (a), (c) K-spaces computed from (b)

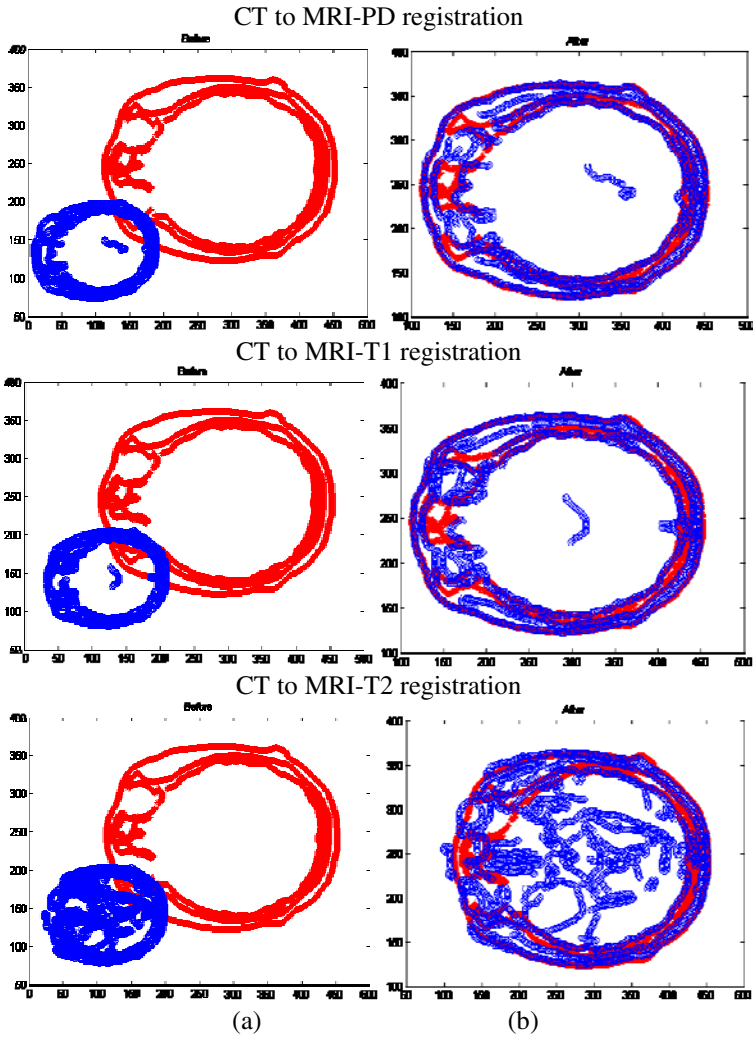


Fig. 2. Example images from an axial plane used in the CT-MR registration experiment from Vanderbilt Database: overlay of the two corresponding edge maps (a) before registration and (b) after registration by using the proposed algorithm

where A is the ground-truth 3-by-3 affine matrix, \hat{A} is the estimated affine matrix, and $\|\cdot\|_F$ denotes the Frobenius norm for a matrix.

Table 2. Average relative errors in the image registration experiment by using the proposed algorithm and SPM [18]

MR T1-T2 registration	The proposed method	SPM8[18, 19]
MR-T1 to MR-T2	0.14	0.26

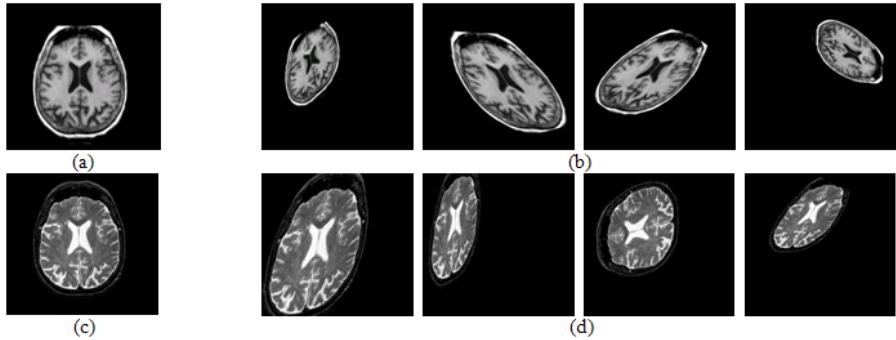


Fig. 3. Example images used in the affine image registration experiment: (a) original T1 brain MRI image, (b) synthesized images from (a) with random affine transforms, (c) original T2 brain MR image, (d) synthesized images from (c) with random affine transforms

5 Conclusions

In this paper, we proposed a robust affine registration algorithm in k-space that can be applied to both image alignment and multi-modal CT-MR registration. The proposed algorithm is based on matching the moments of the Fourier spectrums for two point distributions to alleviate the sensitivity problem of the moment-based approach. To further improve the robustness of the proposed algorithm, we proposed to incorporate the distance weighting to the canny edge for iterative refinement of the affine registration. Our experiments demonstrated the superior performance of the proposed affine image registration algorithm which is suitable for CT-MR registration. In the future, we aim to extend this robust affine registration algorithm to overcome more challenging multi-modal medical image registration problems.

References

1. Hawkes, D.J.: Algorithms for radiological image registration and their clinical application. *Journal of Anatomy* 193(3), 347–361 (1998)
2. XiaoShen, W., LongGen, L., ChaoSu, H., JianJian, Q., ZhiYong, X., Yan, F.: A comparative study of three CT and MRI registration algorithms in nasopharyngeal carcinoma. *Journal of Applied Clinical Medical Physics* 10(2) (2009)
3. Jean-François, D., Mérence, S., Anne, B., Guy, C., Max, L., Vincent, G.: Evaluation of a multimodality image (CT, MRI and PET) coregistration procedure on phantom and head and neck cancer patients: accuracy, reproducibility and consistency. *Radiotherapy & Oncology* 69(3), 237–245 (2003)
4. Antoine Maintz, J.B., Viergever, M.A.: A survey of medical image registration. *Medical Image Analysis* 2(1), 1–36 (1998)
5. Zitová, B., Flusser, J.: Image registration methods: a survey. *Image Vision Computing* 21(11), 977–1000 (2003)

6. West, J., et al.: Comparison and evaluation of retrospective inter-modality brain image registration techniques. *Journal of Computer Assisted Tomography* 21(4), 554–566 (1997)
7. Josien, P.W., Pluim, J.B., Antoine, M., Max, A.V.: Mutual information based registration of medical images: a survey. *IEEE Trans. Med. Imaging* 22(8), 986–1004 (2003)
8. Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., Suetens, P.: Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging* 16(2), 187–198 (1997)
9. Twieg, D.: The k-trajectory formulation of the NMR imaging process with applications in analysis and synthesis of imaging methods. *Medical Physics* 10(5), 610–621 (1983)
10. De Castro, E., Morandi, C.: Registration of translated and rotated images using finite Fourier transforms. *IEEE Trans. Pattern Analysis Mach. Intell.* 3, 700–703 (1987)
11. Reddy, B.S., Chatterji, B.N.: An FFT-based technique for translation, rotation, and scale-invariant image registration. *IEEE Trans. Pattern Analysis Mach. Intell.* 5(8), 1266–1270 (1996)
12. Pan, W., Qin, K., Chen, Y.: An adaptable-multilayer fractional Fourier transform approach for image registration. *IEEE Trans. Pattern Analysis Mach. Intell.* 31(3), 400–413 (2009)
13. Bracewell, R.N., Chang, K.Y., Jha, A.K., Wang, Y.H.: Affine theorem for two-dimensional Fourier transform. *Electronics Letters* 29(3), 304 (1993)
14. Foroosh, H., Zerubia, J.B., Berthod, M.: Extension of phase correlation to subpixel registration. *IEEE Trans. Image Processing* 11(3), 188–200 (2002)
15. Zokai, S., Wolberg, G.: Image registration using log-polar mappings for recovery of large-scale similarity and projective transformations. *IEEE Trans. Image Processing* 14(10), 1422–1434 (2005)
16. Ho, J., Peter, A., Ranganrajan, A., Yang, M.-H.: An algebraic approach to affine registration of point sets. *Proc. Int. Conf. on Computer Vision 2009* (2009)
17. Vanderbilt Database, <http://www.insight-journal.org/rire/index.php>
18. SPM, <http://www.fil.ion.ucl.ac.uk/spm/>
19. D’Agostino, E., Maes, F., Vandermeulen, D., Suetens, P.: Non-rigid atlas-to-image registration by minimization of class-conditional image entropy. In: Barillot, C., Haynor, D.R., Hellier, P. (eds.) *MICCAI 2004, Part I. LNCS*, vol. 3216, pp. 745–753. Springer, Heidelberg (2004)
20. The Whole Brain Atlas, <http://www.med.harvard.edu/AANLIB/home.html>

Sparse Temporal Representations for Facial Expression Recognition

S.W. Chew¹, R. Rana^{1,3}, P. Lucey², S. Lucey³, and S. Sridharan¹

¹ Speech Audio Image and Video Technology Laboratory at Queensland University of Technology, Australia

² Disney Research Pittsburgh

³ Commonwealth Science and Industrial Research Organisation (CSIRO), Australia
{sien.chew,s.sridharan}@qut.edu.au, patrick.lucey@disneyresearch.com,
{rajib.rana,simon.lucey}@csiro.au

Abstract. In automatic facial expression recognition, an increasing number of techniques had been proposed for in the literature that exploits the temporal nature of facial expressions. As all facial expressions are known to evolve over time, it is crucially important for a classifier to be capable of modelling their dynamics. We establish that the method of sparse representation (SR) classifiers proves to be a suitable candidate for this purpose, and subsequently propose a framework for expression dynamics to be efficiently incorporated into its current formulation. We additionally show that for the SR method to be applied effectively, then a certain threshold on image dimensionality must be enforced (unlike in facial recognition problems). Thirdly, we determined that recognition rates may be significantly influenced by the size of the projection matrix Φ . To demonstrate these, a battery of experiments had been conducted on the CK+ dataset for the recognition of the seven prototypic expressions – anger, contempt, disgust, fear, happiness, sadness and surprise – and comparisons have been made between the proposed temporal-SR against the static-SR framework and state-of-the-art support vector machine.

Keywords: sparse representation classification, facial expression recognition, temporal framework

1 Introduction

Advancements made in the field of affective computing research are being rapidly propelled by commercial interests such as marketing, human-computer-interaction, health-care, security, behavioral science, driver safety, etc. A central aim of this research is to enable a computer system to detect the emotional state of a person through various modalities (e.g., face, voice, body, actions), in which the inference through one's facial expression had been a significant contribution. In the recent literature [1,2,3,4], an increasing number of machine learning techniques had been proposed to take advantage of the dynamics inherent in facial expressions. Intuitively, enabling the temporal information of a signal to be exploited serves to elegantly unify a machine learning framework

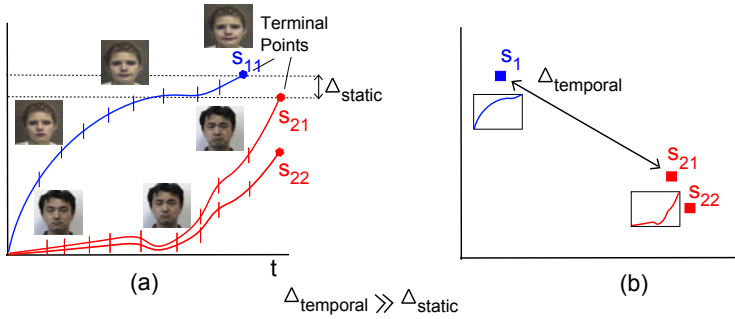


Fig. 1. As a simple thought experiment, consider in (a) a static approach which distinguishes between 2 classes of signals s_1 and s_2 (e.g., happiness vs sadness as the choice of signals here). Observe that only the terminal points in (a) are considered, where the signal's temporal evolution have been discarded (hence small Δ_{static}). On the other hand, a strategy which exploits temporal information is able to map (a) to (b), which enforces both signal classes to occupy a two-dimensional space based on their 'shape' in time (i.e., temporal content); and thus produces a large Δ_{temporal} .

with the architecture of how facial expressions naturally evolve (see Figure 2). To illustrate this concept, observe in Figure 1(a) how the training/prediction of a classifier is determined using only the terminal points (e.g., a single frame containing the expression's apex). A problem with such a static approach is that it takes into account only a single state of the signal, but disregards the signal's past states (i.e., memoryless). By incorporating temporal information (Figure 1(b)), one is able to amplify the minuscule static differences Δ_{static} using the signals' temporal content to obtain a larger Δ_{temporal} . A drawback with adopting such a strategy revolving round a temporal framework, however, is that the complexity of the problem increases proportionally to the quantity of temporal information under consideration (i.e., more training and testing data). Having this in mind, we aim to develop a method which achieves these objectives, while at the same time being able to reduce data dimensionality. One method which gracefully fulfills the latter requirement is the method of sparse representation (SR) classification [5]. However, this method in its current formulation is unable to fulfill the former objective of modelling the dynamics of various expressions. In this paper, we propose a temporal framework for it to fully exploit the dynamics of facial expression signals in an efficient manner (see Section 3).

Recently, the above-mentioned (static) SR classification method had generated considerable excitement in the field of face recognition, thus its transition towards facial expression recognition comes as little surprise. In [5], it was proposed that the downsampling of the input image to a dimensionality of approximately 10×10 pixels, and then projecting this image using a random projection produced impressive performance for the task of person identification. From our experiments, we found that this procedure was not suitable for expression recognition. As opposed to solving for the identity of a person, facial expressions are formed through numerous interactions between various facial muscle groups

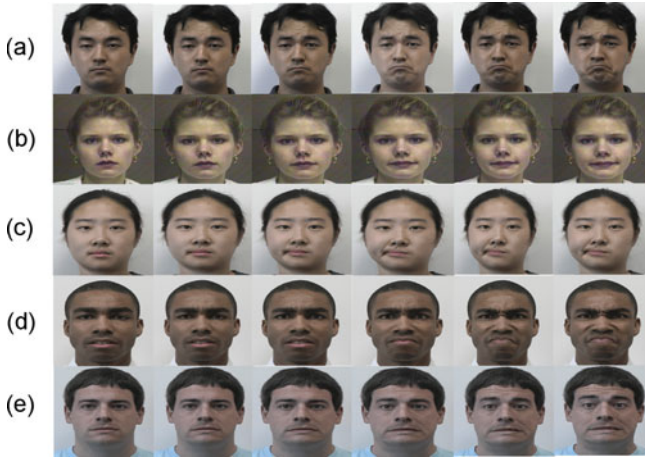


Fig. 2. Examples from the CK+ dataset [10] illustrating the strong temporal links present within neighbouring frames among different expressions, (a) sadness, (b) happiness, (c) contempt, (d) anger and (e) fear

(e.g., eyebrows, lips, nose), most of which require an adequate number of pixels to represent.

In Section 3 we show that the solution to such high dimensionality vectors is computationally exhaustive. We employ SR theory to reduce the image dimensionality, and investigate the impact of different dimensions on recognition performance. Interestingly, different expressions were observed to react differently to this. In most of the works [6,7,8,9] pertaining to using SR classifiers for expression detection, a static approach had been adopted; that is, only single independent frames from various sequences were used. However, it should be recognized that facial expressions are inherently temporal by nature (as shown in Figure 2) and it will be beneficial to incorporate temporal information into the SR classifier.

The central contributions of this paper are,

- Propose a temporal framework for sparse representation classifiers to improve facial expression recognition rates.
- Investigate the effects of downsampling the input images, and the significance of dimensionality reduction on detection accuracy.
- Compare the state-of-the-art SVM framework versus the conventional static-SR classifier and the proposed temporal-SR classifier frameworks and demonstrate that the proposed temporal method offers improved recognition rates. To the best of our knowledge, we are the first to quantitatively report the performance of SR classifiers for all seven expressions. This is important because application developers can decide between SVM and temporal-SR classifier based on the accuracy versus complexity trade-off.

2 Review of Temporal-Based Methods in Expression Recognition

Exploiting temporal information for facial expression recognition is not new. For example, in [4] which used a dynamic Bayesian network to model the dynamic evolution of various facial action units (AUs). A slightly different approach was adopted in [3] which modeled AUs using expression dynamics coupled with phase information. A list of other temporal techniques can be found in [11,12,13]. The major difference between these works and ours, however, is that all these methods proposed require complex features to be extracted from multiple temporal frames. As feature extraction may be considered to be computationally expensive even in the static context, the problem becomes additionally complex and computationally expensive when multiple temporal frames are to be considered. On the other hand, our method does not require any feature representations to be computed. Furthermore, our method is driven by SR theory which is different from all of the above-mentioned methods. SR classification had been used for both face recognition [5] and facial expression recognition [14] previously (especially in [5] which utilized random features). However, none of these SR methods had capitalized on expression dynamics. Our proposed method exploits expression dynamics through SR theory, and we demonstrate the advantages of this method over other feature-based alternatives that rely mainly on only spatial information.

3 A Temporal Framework for Sparse Representation Classification

In this section we describe the underlying mechanics of the proposed Sparse Representation (SR) classifier for facial expression recognition. We initiate the description in terms of static frames, and then introduce the incorporation of temporal information into the framework.

Let us consider that we have N facial expression images spanning over $c = 1, 2, \dots, C$ class. We represent these N images by N vectors $\vec{v}_1, \dots, \vec{v}_N \in \mathbb{R}^n$. Let us construct a dictionary ξ by packing the vectors $\vec{v}_i, \forall i=1, \dots, N$ into the columns of a matrix $\xi \in \mathbb{R}^{n \times N}$. Intuitively, a test sample (e.g., a face image) $\vec{\gamma} \in \mathbb{R}^n$ of class $i \in \{1, 2, \dots, C\}$ can be represented in terms of the dictionary ξ by the following linear combination,

$$\vec{\gamma} = \xi \vec{\alpha}, \quad (1)$$

where $\vec{\alpha} = [0, 0, \pi_{i1}, \dots, \pi_{ik}, 0, 0]^T$, π_{ij} are some scalars and k is the number of face images per class. Clearly, the solution to (1) (i.e. $\vec{\alpha}$) would recognize the test image class (class corresponds to nonzero element is the match), however, we have to identify a method to compute a sparse $\vec{\alpha}$.

A general method to find the sparse solution of (1) is to solve the following optimization problem:

$$\arg \min \hat{\alpha} = \|\vec{\alpha}\|_0, \quad s.t. \vec{\gamma} = \xi \vec{\alpha}, \quad (2)$$

where $\|\cdot\|_0$ denotes the ℓ_0 norm, which returns the nonzero elements of α . Note that if $n > N$, the system is overdetermined, in that case (2) can be solved in polynomial time. Typically the dimension of the image (n) is quite high compared to the available image set, therefore, it is rather impossible for normal computers to solve (2) [5]. For this reason, in practice, the dimension of n is reduced to a smaller size $d \ll n$ (by multiplying a random projection matrix $\Phi \in \mathbb{R}^{d \times n}$ with ξ), which turns (2) into an underdetermined problem. In general, searching for a sparse solution of an underdetermined system using (2) is NP-hard.

Encouragingly, SR theory shows that if $\vec{\alpha}$ is sufficiently sparse, then this underdetermined system can be solved using the following ℓ_1 norm minimization problem, which will produce a similar solution to solving the ℓ_0 norm.

$$\arg \min \vec{\alpha}^* = \|\vec{\alpha}\|_1, \quad s.t. \vec{\gamma} = \xi \vec{\alpha} \tag{3}$$

However, a sparse $\vec{\alpha}$ cannot always guarantee a unique solution to (3). SR theory shows that if $\Theta = \Phi \xi$ obeys the restricted isometry property (RIP) [15], then the underdetermined system (1) can be solved through (3). More encouragingly, SR theory also suggests that Φ obeys RIP. Such as, we use a Φ which is populated by sampling normally distributed numbers with zero mean and variance $\frac{1}{d}$ (i.e., $\Phi \sim \mathcal{N}(0, \frac{1}{d})$). SR theory has shown that $\Phi \sim \mathcal{N}(0, \frac{1}{d})$ obeys RIP when $d \propto K \log(\frac{n}{K})$. Here K is the measure of the sparsity of α . In this paper, instead of seeking to determine an optimal d , we investigated the impact of different values of d on the detection accuracy (see Figure 5).

Transitioning to a temporal framework, the most straightforward approach to incorporate the dynamics of a video sequence (i.e., temporal frames) into a SR classifier would be to simply concatenate consecutive frames into ξ , such that $\tilde{\xi} \in \mathbb{R}^{n \times (N \mapsto \tau)}$, where $\tau = Nt$ and t represents the length of the temporal window. Intuitively, one problem with this approach is that the sparsity of the solution $\vec{\alpha}^*$ is ultimately reduced (by a factor of t) due to the increase in dimensionality of ξ . Following this, γ is then said to be composed of an *additional number of terms* in the linear combination $\xi \vec{\alpha}$ (which may be considered to be noise). In order to circumvent this problem, we postulate that the dimensions of ξ must be maintained to be at N number of columns.

In order to retain the size to N , each column in $\tilde{\xi}$ is formed through the fusion of multiple frames into a single frame that is representative of the temporal information in all t frames; such that $t \mapsto 1$ and $\xi \in \mathbb{R}^{n \times \tau} \mapsto \tilde{\xi} \in \mathbb{R}^{n \times N}$. Another point that needs to be addressed is whether the utilization of sparse feature representations (i.e., a sparse $\tilde{\xi}$) would lead to a better dynamic model. We explored a technique described in [7] which utilized absolute difference images (i.e., $|I_{\text{apex}} - I_{\text{neut}}|$) as sparse feature representations for ξ and γ . Theoretically speaking, an argument may be made that the majority of holistic information in the face is effectively removed once the absolute difference operation is performed. This may be undesirable because facial expressions are essentially holistic in nature (i.e., anger or happiness, etc, occurs in the whole face and are therefore not limited to specific local facial regions). To show this, we incorporated absolute

difference feature representations into $\tilde{\xi}$, and compared the performance with using just pixels,

$$\mathbf{D}_j = \left[|I_{(\text{apex}-t)} - I_{\text{neutral}}|, |I_{(\text{apex}-t+1)} - I_{\text{neutral}}|, \dots, |I_{\text{apex}} - I_{\text{neutral}}| \right]; \quad (4)$$

where $j \in \{1, \dots, \tau\}$

where the columns $\mathbf{D}_j \in \mathbb{R}^{n \times t}$ of $\tilde{\xi} \in \mathbb{R}^{n \times \tau}$ are calculated using absolute difference image operations on the neutral frame with respect to frames 1 to t . Empirical results shown in Figure 4(b) suggested that utilizing these sparse absolute difference features had led to a deterioration in detection accuracy. More details can be found in Section 5.

Understanding this, we focused on using only pixels (i.e., no features) and analyzed various approaches of temporal fusion to form the columns of $\tilde{\xi}$. Simply computing the global average of all t frames in the temporal window would satisfy this criterion. However, from preliminary experiments conducted, we found that one drawback with such an approach was that a *blurring* phenomenon was induced as a result of registration-error/pixel-misalignments that is inherent in all tracked faces. In order to minimize blurring, we employed local averages of several selected facial regions which we had deemed vital in expression recognition (i.e., eyes and mouth, see Figure 3). The pixels residing outside of these two regions were then filled by the remainder of pixels from the apex frame. Further details on this approach are discussed in Section 5.

More sophisticatedly, temporal fusion can be accommodated by familiar methods such as principal component analysis (PCA) and discrete wavelet transforms, etc. To elucidate, the edges in an image may be considered to be the most salient features [16] perceived by the human visual system. Wavelet decomposition may be employed to extract these salient features, and the combination of these features would thus effectively capture the dynamics of the entire window into a single frame. Similarly with PCA, the most salient features are computed using the eigenvalues of the respective covariance matrices. The complete list of all temporal fusion methods employed in this paper is listed as follows – a) local average (AVG), b) gradient pyramid (GRA), c) laplacian pyramid (LAP), d) principal component analysis (PCA), e) discrete wavelet transform (DWT), f) shift invariant discrete wavelet transform (SID), g) morphological difference pyramid (MOD), and h) filter-subtract-decimate pyramid (FSD). Please refer to Figure 3 for an illustration of the fusion process, and also refer to [17] for an excellent description of these image fusion methods. In Section 5 we demonstrate that there is no universal fusion technique that is better for all the facial expressions. It was also reported in [18] that due to the subjective characteristics of the fusion performance evaluation, it is difficult to recommend a method for a given expression. However, in future we wish to investigate why a given fusion method performs better for a given expression.

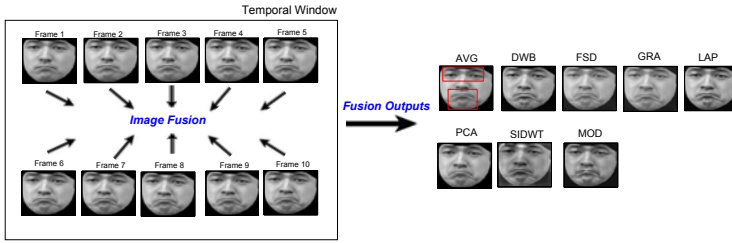


Fig. 3. Fusion of multiple images in a temporal window using i) AVG: local-average (the red bounding boxes illustrate regions where the local averaged were calculated from), ii) DWT: discrete wavelet transform , iii) FSD: filter-subtract-decimate pyramid, iv) GRA: gradient pyramid, v) LAP: laplacian pyramid, vi) PCA: principal component analysis, vii) SIDWT: shift invariant discrete (SID) wavelet transform, and viii) MOD: morphological difference pyramid

4 Experimental Setup

All experiments in this paper had been conducted with the objective of detecting the seven prototypic emotional facial expressions – anger, contempt, disgust, fear, happiness, sadness and surprise – which are available in the CK+ database. Active appearance models (AAMs) were employed for face-tracking, and its corresponding output SAPP pixel representations were used for training and testing the classifiers. For a fair comparison, the exact same two-fold cross validation train/test data partitions were adopted in all evaluations. A subject-independent approach was adopted in all evaluations. We shall adopt μ_{diag} to represent the weighted mean of the diagonal of the confusion matrix as the performance metric in all experiments.

4.1 AAM-Derived Pixel Representations

Active Appearance Models (AAMs) [19] have been shown to be a good method of aligning a pre-defined linear shape model that also has linear appearance variation, to a previously unseen source image containing the object of interest. In general, AAMs fit their shape and appearance components through a gradient-descent search. The shape, $\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^m p_i \mathbf{s}_i$, of an AAM is described by a 2D triangulated mesh, which corresponds to a source appearance image; where $\mathbf{p} = (p_1, \dots, p_m)^T$ are the shape parameters. In all our experiments, we report empirical results obtained from processing AAM-derived similarity normalized appearance features (i.e., SAPP pixel representations).

4.2 The Extended Cohn-Kanade Database

In this paper we used the Extended Cohn-Kanade (CK+) database [10], which contains 593 sequences from 123 subjects. The image sequences vary in duration (from 10 to 60 frames) and incorporate the onset (which is also the neutral frame) to peak formation of the facial expressions. For the 593 posed sequences, full FACS [20] coding of the peak frames had been provided.

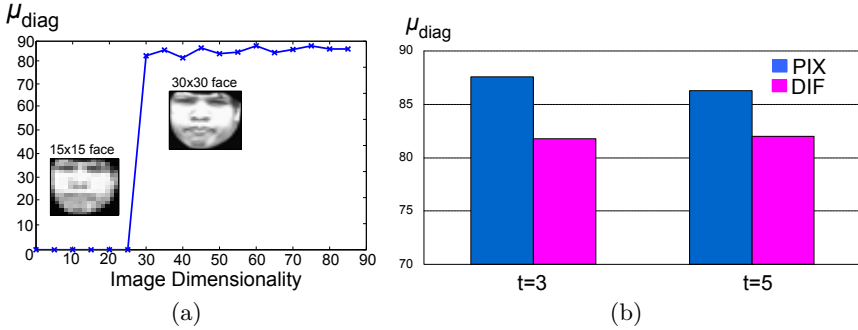


Fig. 4. (a) Once a threshold was exceeded (25×25 pixels), recognition rates were no longer significantly influenced by the image dimensionality. But, if image dimensionality fell below the threshold, then a unique solution to the objective function could not be found. (b) A deterioration in recognition rates was incurred when sparse absolute difference representations (DIF) were utilized in place of raw intensity pixel values (PIX).

5 Experimental Results

As mentioned in Section 3, we wish to first highlight the effect of naively concatenating temporal frames in ξ , such that $\tilde{\xi} \in \mathbb{R}^{n \times (N \rightarrow \tau)}$, is equivalent to the addition of noise; and therefore would have a detrimental effect on recognition rates. We further demonstrate that no benefits are introduced from using difference images due to a *lossy* effect inherent in subtractive operations on the holistic face. In fact, taking absolute difference images had produced substantial deterioration, which was mainly due to holistic information lost from the subtraction. These were supported by empirical results presented in Figure 4(b). In the SR classifier, the dimension of the random projection matrix Φ was taken to be a quarter that of the input image (we shall denote this by $\lambda = \frac{n}{d} = 4$).

5.1 Investigating Temporal Fusion and The Dimension of the Random Projection Matrix

In order to rectify the problems discussed in the previous section, two objectives must be fulfilled: i) the dimension N in ξ should not increase to obtain the sparsest $\bar{\alpha}^*$, and ii) holistic face information must be retained. These two criteria may be easily fulfilled through the utilization of image fusion techniques. All image fusion methods listed in Section 3 had been explored in our experiments. Concerning image dimensionality, we found that downsampling of facial images to a very low dimensionality was not suitable for expression recognition (unlike in face recognition [5]). Figure 4(a) shows that when all other variables except image dimensionality were held constant ($\lambda = 4$ and $t = 1$ in this experiment), the mean detection accuracy was not influenced once a threshold (25×25 pixels) on the image dimensionality was exceeded. In all subsequent experiments, the

original image dimensionality had been preserved (87×93 pixels). It had also been observed that the dimension of the random projection matrix played a significant role in emotional expression detection. Denoting $\lambda = \frac{n}{d}$ as the factor by which the dimension of the projection matrix is downsampled with respect to the dimension of the input image, we were interested in analyzing the effect that varying the temporal window length t and λ had on recognition rates. The 3D plots (Accuracy versus time(t) versus λ) shown in Figure 5 shows that the optimal λ and t can be very different for different expressions. We observed that a larger t was more suitable for contempt, a larger λ was more suitable for sadness, and a larger λ coupled with a mid-range t was more suitable for anger. However, these two variables did not significantly influence happiness, disgust and surprise; which achieved near-perfect detections in our experiments.

5.2 Discussion

Recognition rates of the proposed temporal-SR classifier versus the static-SR classifier is presented in Table 1. For completeness, we have also included the performance of linear SVMs (which was trained and tested on in the exact same

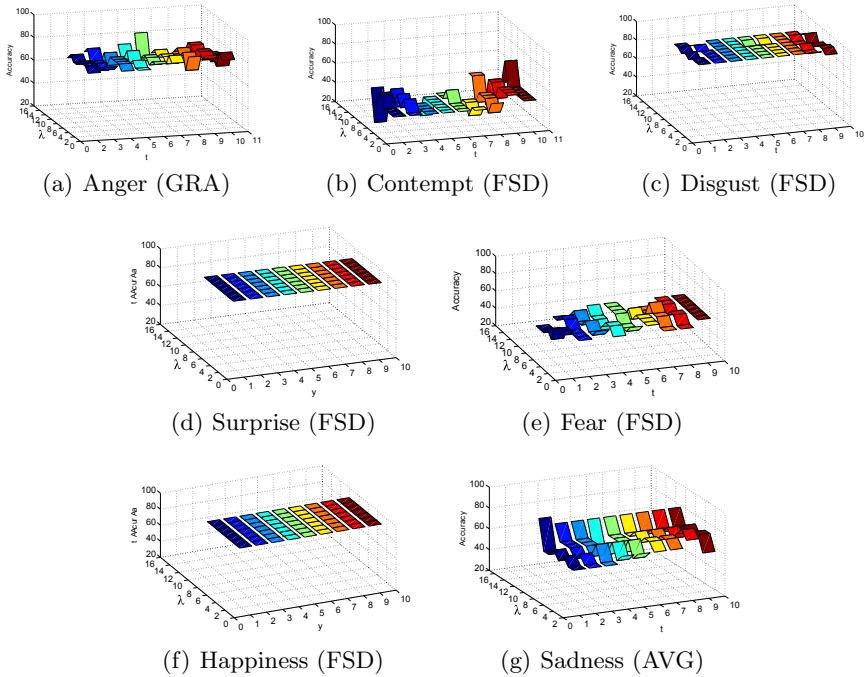


Fig. 5. 3D plots (X-axis (time t), Y-axis (random projection matrix dimension downsampling factor $\lambda = \frac{n}{d}$), Z-axis (Accuracy)) of the detection performances of the seven emotions as functions of time and λ . The image fusion method which produced the best recognition accuracy is shown in brackets.

Table 1. Recognition rates for emotion classification on the CK+ dataset for static-SR versus temporal-SR classification, and referenced to a linear SVM. μ_{diag} represents the weighted mean of the diagonal of the respective confusion matrices (computed through majority voting), and \mathbf{N} represents the number of examples available from each emotion.

	\mathbf{N}	static-SRC	temporal-SRC	SVM
Anger	45	90.9	95.5	86.1
Contempt	18	55.6	75.0	55.6
Disgust	59	100.0	100.0	100.0
Fear	25	66.7	75.0	91.7
Happiness	69	100.0	100.0	100.0
Sadness	28	85.7	92.9	85.7
Surprise	83	97.6	97.6	97.6
μ_{diag}	—	91.9	94.9	93.2

manner). As can be seen, the static-SR method experienced a deterioration of 1.3% with respect to the SVM, but once temporal information had been incorporated into the SR classifier, then a 3% improvement of the temporal-SR method over the static-SR method was afforded. Although it may appear on the surface that the differences between all three methods are not very significant, but we should not ignore the fact that the asymptote of ideal detection (i.e., perfect 100% recognition) is being approached and slight differences of a few percent may be more significant than as it would appear.

In view of this, it would be profoundly more interesting for an investigation to be conducted on more realistic facial expressions (i.e., acted and spontaneous) which possess deeper temporal dependencies for further insights of the underlying mechanisms of both static- and temporal-SR classifiers to be gained. In addition, since SVMs have been actively employed in expression recognition, it would also be interesting to make a direct comparison with the SR classifiers by employing fused temporal information. Such an analysis would stimulate an interesting thought-provoking analysis on which is more capable at exploiting expression dynamics – the ℓ_1 -norm or the ℓ_2 -norm?

6 Conclusion and Future Work

In this paper, we explored the method of sparse representation (SR) classification to detect the seven prototypic emotion-related facial expressions. Having established the importance of expression dynamics, we proposed a framework in which a dynamic model could be effectively implemented into the SR classifier. Our work explored the logic behind the use of sparse features in the SR framework, and also investigated the influence of the dimensions of the random projection matrix and length of the temporal window. Indeed, we found that the latter two were significant factors in influencing detection performance. Additionally, various techniques of incorporating temporal information into feature

matrix ξ had been analyzed and proposed. In future work, we intend to investigate if the dynamics of more realistic and spontaneous facial expressions (on both emotional-related expressions and action units) could be exploited using our proposed method. Apart from this, we wish to analyze in further detail why the filter-subtract-decimate pyramid image fusion method was more suitable for most expressions, but not for the remaining few.

Acknowledgments. This research was supported in part by the Cooperative Research Centre for Advanced Automotive Technology (AutoCRC).

References

1. Kobayashi, H., Hara, F., Ikeda, S., Yamada, H.: A basic study of dynamic recognition of human facial expressions. In: 2nd IEEE International Workshop on Robot and Human Communication, pp. 271–275 (1993)
2. Pantic, M., Patras, I.: Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. In: IEEE International Conference on Systems, Man and Cybernetics, vol. 4, pp. 3358–3363 (2005)
3. Valstar, M., Pantic, M.: Fully automatic facial action unit detection and temporal analysis. In: Computer Vision and Pattern Recognition Workshop CVPRW 2006, pp. 149–149 (2006)
4. Tong, Y., Liao, W., Ji, Q.: Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 1683–1699 (2007)
5. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2008)
6. Bociu, I., Pitas, I.: A new sparse image representation algorithm applied to facial expression recognition. In: Proceedings of the 14th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing, 2004, p. 539 (2004)
7. Zafeiriou, S., Petrou, M.: Sparse representations for facial expressions recognition via l_1 optimization. In: Computer Vision and Pattern Recognition Workshops (CVPRW), p. 32 (2010)
8. Cotter, S.: Recognition of occluded facial expressions using a fusion of localized sparse representation classifiers. In: Digital Signal Processing Workshop and IEEE Signal Processing Education Workshop (DSP/SPE), p. 437 (2011)
9. Mahoor, M., Zhou, M., Veon, K.L., Mavadati, S., Cohn, J.: Facial action unit recognition with sparse representation. In: Automatic Face & Gesture Recognition and Workshops, p. 336 (2011)
10. Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: Proceedings of the IEEE Workshop on CVPR for Human Communicative Behavior Analysis (2010)
11. Yang, P., Liu, Q., Cui, X., Metaxas, D.: Facial expression recognition using encoded dynamic features. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
12. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29, 915–928 (2007)

13. Du, R., Wu, Q., He, X., Jia, W., Wei, D.: Facial expression recognition using histogram variances faces. In: 2009 Workshop on Applications of Computer Vision, WACV (2009)
14. Ying, Z.-L., Wang, Z.-W., Huang, M.-W.: Facial Expression Recognition Based on Fusion of Sparse Representation. In: Huang, D.-S., Zhang, X., Reyes García, C.A., Zhang, L. (eds.) ICIC 2010. LNCS, vol. 6216, pp. 457–464. Springer, Heidelberg (2010)
15. Candes, E.: The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathématique* 346, 589–592 (2008)
16. Hubel, D.: *Eye, Brain and Vision*. Freeman (1987)
17. Rockinger, O., Fechner, T.: Pixel-level image fusion: The case of image sequences. In: *Proc. SPIE*, vol. 3374, pp. 378–388 (1998)
18. Canga, E.F.: Image fusion. Project report, Dept. Electronic and Electrical Eng., Univ. of Bath. (2002)
19. Cootes, T., Edwards, G., Taylor, C.: Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 681–685 (2001)
20. Ekman, P.: *Emotion in the human face*. Cambridge University Press (1982)

Dynamic Compression of Curve-Based Point Cloud

Ismael Daribo¹, Ryo Furukawa¹, Ryusuke Sagawa², Hiroshi Kawasaki³,
Shinsaku Hiura¹, and Naoki Asada¹

¹ Faculty of Information Sciences, Hiroshima City University, Hiroshima, Japan

² National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan

³ Faculty of Engineering, Kagoshima University, Kagoshima, Japan
{daribo,ryo-f,hiura,asada}@hiroshima-cu.ac.jp,
ryusuke.sagawa@aist.go.jp, kawasaki@ibe.kagoshima-u.ac.jp

Abstract. With the increasing demands for highly detailed 3D data, dynamic scanning systems are capable of producing 3D+t (*a.k.a.* 4D) spatio-temporal models with millions of points recently. As a consequence, effective 4D geometry compression schemes are required to face the need to store/transmit the huge amount of data, in addition to classical static 3D data. In this paper, we propose a 4D spatio-temporal point cloud encoder via a curve-based representation of the point cloud, particularly well-suited for dynamic structured-light-based scanning systems, wherein a grid pattern is projected onto the surface object. The object surface is then naturally sampled in a series of curves, due to the grid pattern. This motivates our choice to leverage a curve-based representation to remove the spatial and temporal correlation of the sampled point along the scanning directions through a competitive-based predictive encoder that includes different spatio-temporal prediction modes. Experimental results show the significant gain obtained with the proposed method.

Keywords: Point cloud, compression, curve-based, dynamic, 4D, 3D+t, grid pattern.

1 Introduction

Recent evolutions in acquisition technologies allow to produce 3D geometric models with millions of points that evolve over time (see Fig. 1). Problems of efficiently storing, transmitting, processing and rendering spatio-temporal data are then been raised, in addition to classical static 3D data. Towards an efficient compression performance, a suitable 4D data representation becomes particularly more and more important.

Currently active 3D scanners are widely used for acquiring 3D models [2]. Especially, scanning systems based on structured light have been intensively studied in the acquisition of dynamic scene [1,3,4], recently. Structured-light-based scanning is done by sampling the surface of an object with a known pattern (*e.g.* grid, horizontal bars, lines). Studying the deformation of the pattern

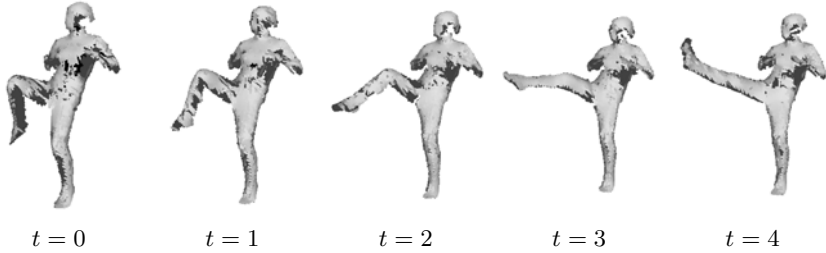


Fig. 1. Example of a 4D data acquired with a grid-pattern one-shot scanners [1]

allows building a 3D model by means of a point cloud. The huge amount of raw point data has to be stored/transmitted by efficient compact means. Noise and incompleteness, however, make the process more difficult to achieve. While mesh compression is a mature field [5], there is still room for improvement in point cloud compression. To the best of our knowledge, present point-based compression strategies are mainly based on surface approximation [6,7], and/or hierarchical space subdivision by augmenting the dataset by a data structure (*e.g.* spanning tree [8,9], octree [10,11]), which lead to either smooth out sharp features, an extra-transmission of a data structure, or an unavoidable lossy encoding. In addition, the augmentation of the dataset by data structure makes difficult the exploitation of temporal consistency.

With the aim of tackling the aforementioned issues, we first made the observation that structured-light-based scanning systems output points along the measuring direction, which naturally orders groups of points along the same direction: scan lines. We particularly aim at structured-light-based scanning systems that use a grid pattern formed by straight lines distinguishable only as vertical and horizontal lines [1] as illustrated in Fig. 2. When the projected grid pattern is extracted from the captured image, 3D points are naturally fitted into a series of space curves. This motivates our choice to leverage the spatially sequential order of the sampled-points along these scan lines: first, we pre-process the data to retrieve each scan line into a curve of points, and after, we exploit the curve-based representation through a spatio-temporal competition-based predictive encoder specially designed with linear and curved-driven prediction modes. We then formulate the problem of encoding a point cloud as “How to retrieve the scan lines?”, and after “How to encode a curve in space?”, and then “How to encode a space curve over time?”. This formulation has the benefit to simplify the former problem into sub-problems that allow application-oriented functionalities, such as:

- temporal prediction that consists of searching for a similar curve in previous temporal frames,
- random access that allows the decoding of a local part of the point cloud without the necessity of decoding the full dataset,
- error propagation limitation since all curves within a frame are independently encoded,

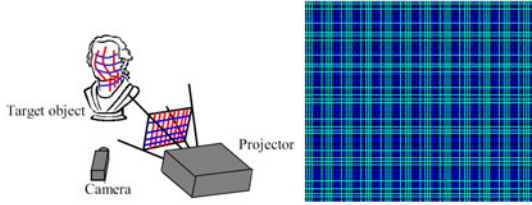


Fig. 2. (left) Grid-pattern-based scanning system: a grid pattern is projected from the projector and captured by the camera. (right) Example of projected grid pattern.

- possible lossless coding due to the predictive nature of the encoder,
- parallel computation where each curve is simultaneously encoded.

The rest of the paper is organized as follows. Section 2 describes the point cloud pre-processing towards a curve-based representation. Section 3 addresses the problem of efficiently compressing a raw point dataset, followed by the experimental results in Section 4. Finally, our final conclusions are drawn in Section 5.

2 Curve-Based Representation

In an arbitrary point cloud, the identification of neighbors is a nontrivial task. One approach consist in locally defining as neighbor the point that minimizes an error cost functional based on a prediction rule, which results in the augmentation of the dataset by a predictive data structure such as a spanning tree that also needs to be encoded and transmitted. However, for point cloud outputted by scanners using structured light there is a straightforward way through the scan lines as discussed before. Moreover, the projected grid-pattern makes each scan lines differentiable from others as shown in Fig. 3. It is then possible to address this problem in a more global way, at least at a scan line scale. In some cases, space curves can be directly obtained from the acquisition process, *e.g.* line detection algorithm [4].

2.1 Curve-Based Point Cloud Definition

Let us consider the point cloud $\mathcal{S} = \{p_1, p_2, \dots, p_N\}$ as a collection of N 3D points $p_{k_{1 \leq k \leq N}}$. As mentioned earlier, structured-light-based 3D scanning systems fit the sampled points in curves. The point cloud \mathcal{S} can then be represented as a set of M curves $\mathcal{C}^{l_{1 \leq l \leq M}}$ as

$$\mathcal{S} = \{\mathcal{C}^1, \mathcal{C}^2, \dots, \mathcal{C}^M\} \tag{1}$$

where a l-ieme curve \mathcal{C}^l is expressed as

$$\mathcal{C}^l = \{p_r, p_{r+1}, \dots, p_s\} \text{ with } 1 \leq r < s < N \tag{2}$$

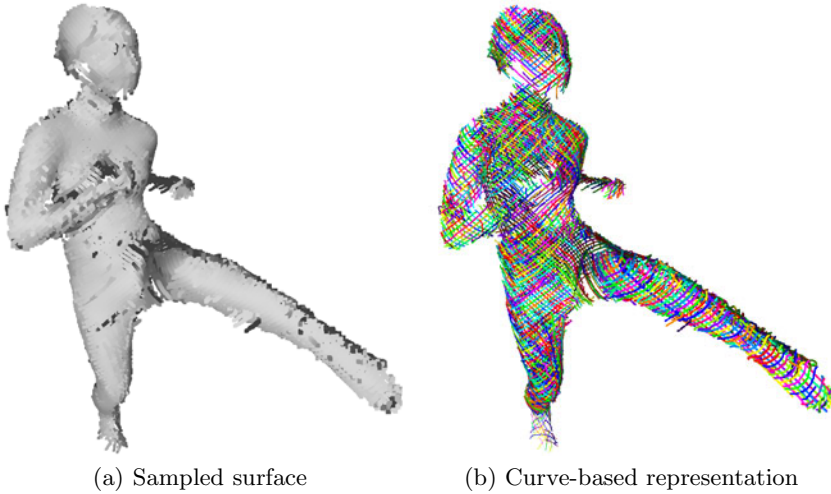


Fig. 3. Sampled point cloud partitioned into a series of curves wrt to the scanning directions. Curves are discriminate by different colors.

2.2 Curve-Based Partitioning

Each curve \mathcal{C} is defined to contain points that share similar proprieties, *e.g.* curvature, direction, Euclidean distance with his neighbor, *etc.*. Algorithm 1 shows how the point cloud \mathcal{S} is partitioned into a set of curves as defined in Equation (1). The division is controlled by defining if the current point p_k to process is an outlier with respect to the current curve \mathcal{C} . In this study, we defined an outlier as

$$d(p_k, p_{k-1}) > \epsilon, \quad (3)$$

$$\text{with } \epsilon = \frac{1}{N-1} \sum_{i=2}^N d(p_i, p_{i-1}).$$

The current point p_k is considered as an outlier and then added to a new curve, if the Euclidean distance $d(.,.)$ is larger than a defined threshold ϵ : here the average value of the distance between two consecutive points throughout the point cloud. In general, other outlier definitions can be considered. For example by checking if adding the current point p_k will disturb the normal distribution of the current curve. Another example is the use of a multiple of the inter-quartile-range (IQR) value of the current curve as a threshold. In this study we consider the Euclidean distance that gives a satisfactory curve set representation.

3 Point Cloud Encoding

After pre-processing the point cloud data to leverage the spatial order of samples along the scanning directions, we propose applying well-known

Algorithm 1. The partitioning algorithm

```

Input: set of points  $\mathcal{S}^*$ 
Output: set of curves  $\mathcal{S}$ 
Data: current curve  $\mathcal{C}$ 

foreach point  $p_k$  in  $\mathcal{S}^*$  do
  if size ( $\mathcal{C}$ ) is lower than 3 then
    | add  $p_k$  in  $\mathcal{C}$ ;
  else if true == isAnOutlier ( $\mathcal{C}, p_k$ ) then
    | add  $\mathcal{C}$  into  $\mathcal{S}$ ;
    | clear  $\mathcal{C}$ ;
    | add  $p_k$  in  $\mathcal{C}$ ;
  else
    | add  $p_k$  into  $\mathcal{C}$ ;
  end
end
if true == isEmpty ( $\mathcal{C}$ ) then
  | add  $\mathcal{C}$  into  $\mathcal{S}$ ;
end

```

hybrid-video-encoding techniques: spatio-temporal predictions followed by residual coding. In opposition to present predictive strategies [12,8,9] wherein a spanning tree is built, prior to only one predictive rule, we propose using multiple spatio-temporal predictors that allow sidestepping the utilization of a spanning tree, and moreover, providing an easier and intuitive way to perform temporal prediction. The proposed framework can be denoted as a competition-based predictive encoder in relation with the competition between all spatio-temporal prediction modes for each point.

Let us first introduce some notations. Let \mathcal{C}^t be the current curve to encode at time t . Each point p_k^t in \mathcal{C}^t is predicted by the prediction \widehat{p}_k^t with respect to the previous coded point \widetilde{p}_i^j with $j \leq t$ and $i < k$. Note that previous coded points have been quantized and inverse quantized. The prediction unit outputs then the quantized corrective vector $r_k^t = p_k^t - \widehat{p}_k^t$, also denoted as residual, and transmits it to the entropy coder. The coding efficiency comes with the accuracy of the prediction that is improved by choosing the most suitable prediction method for each point. The prediction that minimizes the quantized Euclidean distance $\mathcal{Q}(\|p_k - \widehat{p}_k^t\|)$ is defined as the best one. \mathcal{Q} being the quantization operator defined in Section 3.3. Then for each point the chosen prediction mode is signaled in the bitstream. The following predictions obey the two assumptions: closed points within a curve either evolve in the same direction, or turn constantly.

3.1 Intra-prediction

Intra-prediction attempts to determine, for each point p_k in \mathcal{C} , the best predicted point \widehat{p}_k with respect to the previous coded points $\widetilde{p}_{i, i < k}$ in the same curve \mathcal{C} . For notation concision, let us define the sub-curve containing the previous coded points by

$$\mathcal{C}|_{i < k} = \mathcal{C} \cap \{p_i | i < k\}, \quad (4)$$

and the prediction by

$$\hat{p}_k = P(\mathcal{C}|_{i < k}). \quad (5)$$

Note that in this section, we have not made explicit that \mathcal{C} and \hat{p}_k are function of t for notation concision.

No-prediction P^{Intra} . When no-prediction is applied, this defines the current point as a key point used for random access and error propagation limitation.

$$P^{Intra}(\mathcal{C}|_{i < k}) = (0, 0, 0). \quad (6)$$

Const P^{Const} . The previous coded point p_{k-1} in the curve is used as predictor.

$$P^{Const}(\mathcal{C}|_{i < k}) = \tilde{p}_{k-1}. \quad (7)$$

Linear P^{Linear} . The prediction is based on the two previous coded point p_{k-1} and p_{k-2} in the curve, assuming then that p_{k-2} , p_{k-1} and p_k belong to the same line.

$$P^{Linear}(\mathcal{C}|_{i < k}) = 2 \cdot \tilde{p}_{k-1} - \tilde{p}_{k-2} \quad (8)$$

Fit-a-sub-line $P^{FitSubLine}$. The prediction point is an extension of a segment $\mathcal{L}(\mathcal{C}|_{i_0 \leq i < k})$. The segment is given by line fitting algorithm based on the M-estimator technique, that iteratively fits the segment using the weighted least-squares algorithm. The starting point p_{i_0} has to be signaled to the decoder, and thus, an additional flag is put in the bitstream.

$$P^{FitSubLine}(\mathcal{C}|_{i < k}) = 2 \cdot \langle \mathcal{L}(\mathcal{C}|_{i_0 \leq i < k}) \perp \tilde{p}_{k-1} \rangle - \langle \mathcal{L}(\mathcal{C}|_{i_0 \leq i < k}) \perp \tilde{p}_{k-2} \rangle \quad (9)$$

Turning-angle $P^{Turning}$. The current point p_k is predicted under the assumption that the curve is turning constantly around the point p_{k-1} . Given the displacement vector $\mathbf{v}_k = p_k - p_{k-1}$ between two consecutive points, the assumption is equivalent to consider equal the turning angles between two consecutive displacement vectors. The turning angle in 3D space between two vectors being defined by the triplet angles $\alpha(\alpha_x, \alpha_y, \alpha_z)$ as the difference of their direction angles. Given a vector $\mathbf{v}_k(v_{kx}, v_{ky}, v_{kz})$, his direction angles $\theta(\theta_{v_{kx}}, \theta_{v_{ky}}, \theta_{v_{kz}})$ are expressed by

$$\cos(\theta_{v_{kx}}) = \frac{v_{kx}}{\|\mathbf{v}_k\|}, \quad \cos(\theta_{v_{ky}}) = \frac{v_{ky}}{\|\mathbf{v}_k\|}, \quad \cos(\theta_{v_{kz}}) = \frac{v_{kz}}{\|\mathbf{v}_k\|}. \quad (10)$$

The prediction of the current point p_k can then be expressed as

$$P^{Turning}(\mathcal{C}|_{i < k}) = \tilde{p}_{k-1} + \mathcal{R}(\alpha_{k-1}) \cdot \mathbf{v}_{k-1}, \quad (11)$$

where $\mathcal{R}(\alpha_{k-1}) \cdot \mathbf{v}_{k-1}$ being the 3D rotation of \mathbf{v}_{k-1} wrt the turning angle

$$\alpha_{k-1} = (\theta_{v_{k-1}} - \theta_{v_{k-2}}). \quad (12)$$

3.2 Temporal-Prediction

The temporal prediction can be decomposed in two steps: first, find a close curve in previous frames in terms of distance or similarity, and after, use either the direction or curvature of the optimal curve \mathcal{C}^{t-1} to predict the current point p_k^t in \mathcal{C}^t . We then highlight the two main strategies:

- find the closest curve in previous frames by segment matching,
- find the most similar curve with closest Euclidean invariant signature in previous frames,

which result in temporal prediction modes presented thereafter.

Close P^{Close} . Between two consecutive frames, points occupy roughly the same position due to the small amount of motion, or the presence of static objects in the scene. It is then possible to find an optimal curve $\widehat{\mathcal{C}}^{t-1}$ in the previous frame that can be superposed with the current one \mathcal{C}^t such that

$$\widehat{\mathcal{C}}^{t-1} = \underset{\mathcal{C}^{t-1}}{\operatorname{argmin}} \left\{ d(\mathcal{C}^{t-1}|_{i_0 \leq i < k}, \mathcal{C}^t|_{i_0 \leq i < k}) \right\} \tag{13}$$

where the distance between the two curves is expressed by

$$d(\mathcal{C}^{t-1}|_{i_0 \leq i < k}, \mathcal{C}^t|_{i_0 \leq i < k}) = \sum_{i_0}^{k-1} d(\widehat{p}_i^{t-1}, \widehat{p}_i^t). \tag{14}$$

The points in $\widehat{\mathcal{C}}^{t-1}$ are then utilized to predict the current point as follows:

$$P^{Close}(\mathcal{C}^t|_{i < k}) = \widehat{p}_k^{t-1} \tag{15}$$

For instance, other alternatives may consist in considering instead \widehat{p}_{k-1}^{t-1} or the mean point $\frac{1}{2}(\widehat{p}_{k-1}^{t-1} + \widehat{p}_k^{t-1})$.

Similar $P^{Similar}$. By finding a similar curve in terms of shape prior to a close Euclidean invariant signature, we assign the same turning angle, as defined previously, to the predictor. Let us first define the signature of a space curve, up to a Euclidean transform, by its curvature function $\kappa(n)$ and torsion function $\tau(n)$, both functions of the parameter n . It was shown in [13,14] that $\kappa(n)$ and $\tau(n)$ can be approximated at the point p_k by

$$\kappa(p_k) = \pm 4 \cdot \frac{\sqrt{s \cdot (s-a) \cdot (s-b) \cdot (s-c)}}{a \cdot b \cdot c}, \tag{16}$$

$$\tau(p_k) = \pm 6 \cdot \frac{H}{d \cdot e \cdot f \cdot \kappa(p_k)}. \tag{17}$$

where H being the height of the tetrahedron form by $p_{i-1}, p_i, p_{i+1}, p_{i+2}$ of base p_{i-1}, p_i, p_{i+1} , and

$$\begin{aligned} a &= d(p_{k-1}, p_k), & b &= d(p_k, p_{k+1}), & c &= d(p_{k-1}, p_{k+1}), \\ d &= d(p_{k+i}, p_{k+2}), & e &= d(p_k, p_{k+2}), & f &= d(p_{k-1}, p_{k+2}), \end{aligned}$$

and $s = \frac{1}{2}(a + b + c)$. Since κ and τ only depends on the Euclidean distance $d(.,.)$ between points, they provide a completely Euclidean invariant numerical signature approximation. The turning angles of the curve having the closest signature is then applied such that

$$P^{Similar}(\mathcal{C}^t|_{i < k}) = \tilde{p}_{k-1}^t + \mathcal{R}(\alpha_k^{t-1}) \cdot \mathbf{v}_{k-1}^t, \quad (18)$$

where

$$\alpha_k^{t-1} = \left(\theta_{v_k^{t-1}} - \theta_{v_{k-1}^{t-1}} \right). \quad (19)$$

3.3 Quantization

After prediction, the point cloud is represented by a set of corrective vectors, wherein each coordinate is a real floating number. The quantization will enable the mapping of these continuous set of values to a relatively small discrete and finite set. In that sense, we apply a scalar quantization as follow

$$\mathcal{Q}(r_k) = \tilde{r}_k = \text{sign}(r_k) \cdot \text{round}(|r_k| * 2^{bp-1}) \quad (20)$$

where bp is the desired bit precision to represent the absolute floating value of the residual.

3.4 Coding

The last stage of the encoding process removes the statistical redundancy in the quantized absolute component of the residual $|\tilde{r}_k|$ by entropy Huffman coding. Huffman coding assigns a variable length code to each absolute value of the quantization residual based on the probability of occurrence. The bitstream consists of: a header containing the canonical Huffman codeword lengths, the quantization parameter bp , the total number of points and the residual data for every point; wherein the coded residual of every point is composed of: 3 bits signaling the prediction used, 1 bit for the sign, a variable-length code for each absolute component value of the corrective vector with regards to the entropy coder.

4 Experimental Results

The performance of the proposed framework is evaluated using the two models shown in Fig. 4 over 24 frames. We defined a group of frames consisting of twelve frames wherein the first frame is intra-only predicted to limit temporal artifact propagation. The objective compression performance of the proposed method is investigated in the rate-distortion (RD) curves plotted in Figure 5 through the average number of bits per points (bpp), in relation to the loss of quality, measured by the peak signal to noise ratio (PSNR). The PSNR is evaluated using the Euclidean distance between points. The peak signal is given by the length of the diagonal of the bounding box of the original model. The RD



(a) Kick



(b) Squirt

Fig. 4. Test models

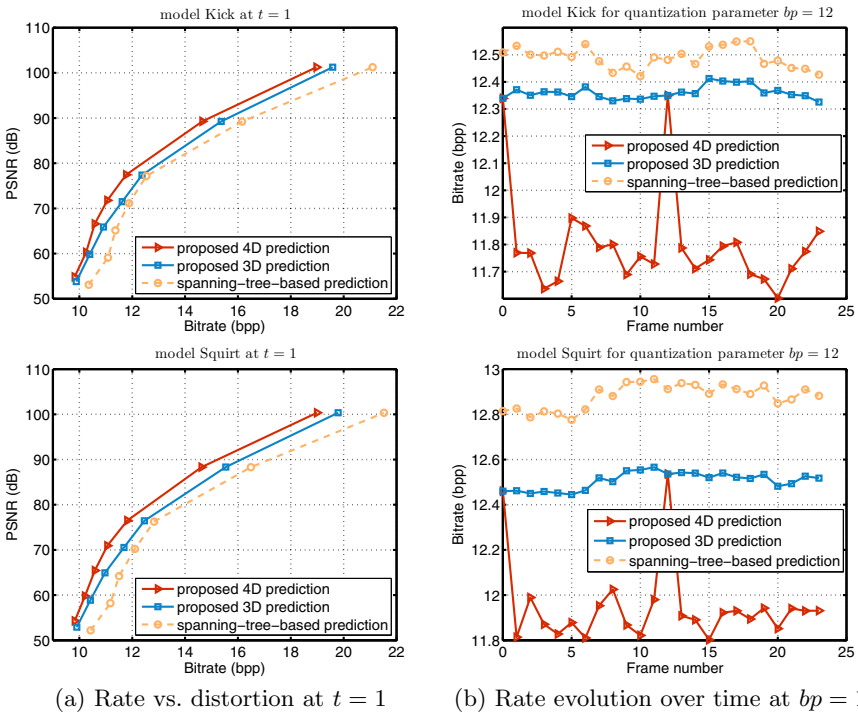
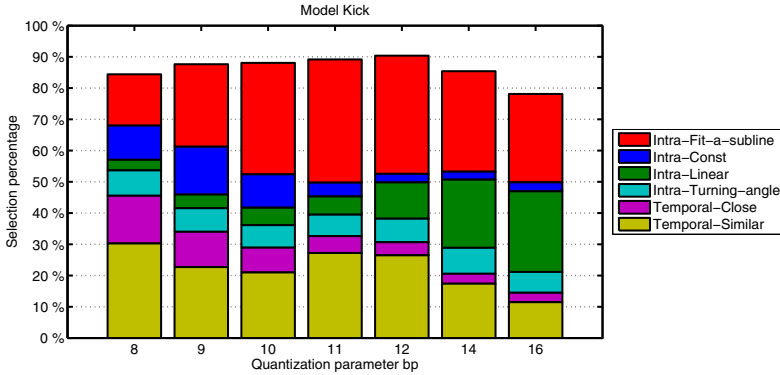
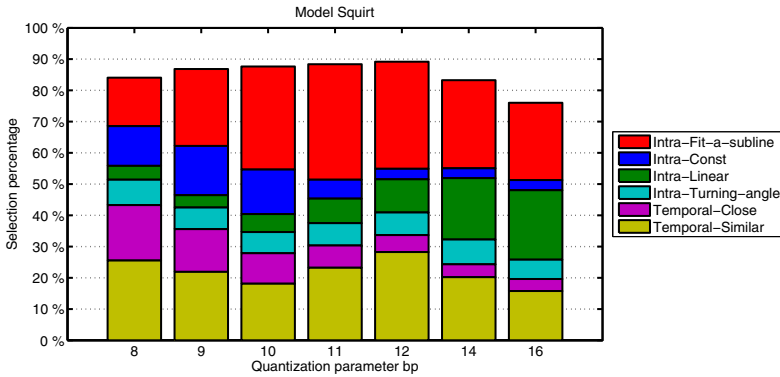


Fig. 5. Rate-distortion performance of the proposed encoder, using 4D and 3D prediction strategy. For comparison, we also show results for classical spanning-tree-based encoder [8]. Experiments are done for 24 frames, where frame at $t = 0$ and $t = 12$ are intra-only encoded.

results correspond respectively to the seven bp quantization parameters: 8, 9, 10, 11, 12, 14 and 16. We compare our intra-only (*a.k.a.* 3D) competitive-optimized strategy and its temporal extension (*a.k.a.* 4D) with the spanning-tree-based strategy [8]. It can be observed that the proposed method provides better RD results in both cases. Experimental results highlight the advantage of competing



(a)



(b)

Fig. 6. Example of average prediction mode distribution at different quantization parameters bp

multiple predictors instead of the spanning tree strategy optimized for only one predictor. It can be observed in Fig. 6 the average distribution of the different prediction modes at different quantization parameter bp . It is important to note that within the prediction unit, the utilized previous coded points have been quantized and inverse quantized, which results in more or less quantization error wrt the quantization parameter bp . Fig. 6 illustrates that at strong quantization (*i.e.* low bp) temporal predictions are more efficient, while at weak quantization intra-linear prediction are over utilized. Ideally, the choice of the prediction mode should be optimized in a rate-distortion sense, which will be left as future work.

5 Conclusion

We designed and implemented a 3D+t (*a.k.a.* 4D) predictive single-rate encoder for point positions outputted by structured-light 3D scanning systems using a grid pattern. We pre-processed the point cloud to leverage points along a series of curves to exploit the spatio-temporal organization of the points, which are ordered prior to the scanning direction. By using curves as a coding unit we succeed in improving the rate-distortion performance, while enabling application-orientated features such as random access, error propagation limitation. Several issues remain that warrant further research. In future studies, we integrate other point attributes (*e.g.* color, normal, *etc.*), and extend our encoder to arbitrary point clouds.

References

1. Furukawa, R., Sagawa, R., Kawasaki, H., Sakashita, K., Yagi, Y., Asada, N.: One-shot entire shape acquisition method using multiple projectors and cameras. In: Proc. of the Pacific-Rim Symposium Image and Video Technology (PSIVT), Singapore, pp. 107–114 (December 2010)
2. Batlle, J., Mouaddib, E.M., Salvi, J.: Recent progress in coded structured light as a technique to solve the correspondence problem: a survey. *Pattern Recognition* 31, 963–982 (1998)
3. Kawasaki, H., Furukawa, R., Sagawa, R., Yagi, Y.: Dynamic scene shape reconstruction using a single structured light pattern. In: Proc. of the IEEE Computer Vision and Pattern Recognition (CVPR), Anchorage, Alaska, pp. 1–8 (June 2008)
4. Sagawa, R., Ota, Y., Yagi, Y., Furukawa, R., Asada, N., Kawasaki, H.: Dense 3D reconstruction method using a single pattern for fast moving object. In: Proc. of the IEEE International Conference on Computer Vision (ICCV), Kyoto, Japan, pp. 1779–1786 (September 2009)
5. Alliez, P., Gotsman, C.: Recent advances in compression of 3D meshes. In: Proc. of the Advances in Multiresolution for Geometric Modelling, pp. 3–26 (2003)
6. Alexa, M., Behr, J., Cohen-Or, D., Fleishman, S., Levin, D., Silva, C.T.: Computing and rendering point set surfaces. *IEEE Transactions on Visualization and Computer Graphics* 9(1), 3–15 (2003)
7. Hubo, E., Mertens, T., Haber, T., Bekaert, P.: Self-similarity-based compression of point clouds, with application to ray tracing. In: Proc. of the IEEE Eurographics Symposium on Point-Based Graphics, pp. 129–137. Eurographics Association (2007)
8. Gumhold, S., Kami, Z., Isenburg, M., Seidel, H.-P.: Predictive point-cloud compression. In: Proc. of the Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH). ACM, USA (2005)
9. Merry, B., Marais, P., Gain, J.: Compression of dense and regular point clouds. In: Proc. of the Computer graphics, Virtual Reality, Visualisation and Interaction in Africa (AFRIGRAPH), pp. 15–20. ACM (2006)

10. Peng, J., Kuo, C.-C.J.: Geometry-guided progressive lossless 3d mesh coding with octree (OT) decomposition. *ACM Transaction on Graphics* 24, 609–616 (2005), <http://doi.acm.org/10.1145/1073204.1073237>
11. Schnabel, R., Klein, R.: Octree-based point-cloud compression. In: Botsch, M., Chen, B. (eds.) *Proc. of the IEEE Eurographics Symposium on Point-Based Graphics*. Eurographics (July 2006)
12. Taubin, G., Rossignac, J.: Geometric compression through topological surgery. *ACM Transaction on Graphics* 17, 84–115 (1998)
13. Calabi, E., Olver, P., Tannenbaum, C.S.A., Haker, S.: Differential and numerically invariant signature curves applied to object recognition. *International Journal of Computer Vision* 26, 107–135 (1998)
14. Boutin, M.: Numerically invariant signature curves. *International Journal of Computer Vision* 40(3), 235–248 (2000)

Recovering Depth Map from Video with Moving Objects

Hsiao-Wei Chen and Shang-Hong Lai

Computer Science, National Tsing Hua University,
No. 101, Section 2, Kuang-Fu Road, Hsinchu, Taiwan 30013, R.O.C.
wei@hotmail.com, lai@cs.nthu.edu.tw

Abstract. In this paper, we propose a novel approach to reconstructing depth map from a video sequence, which not only considers geometry coherence but also temporal coherence. Most of the previous methods of reconstructing depth map from video are based on the assumption of rigid motion, thus they cannot provide satisfactory depth estimation for regions with moving objects. In this work, we develop a depth estimation algorithm that detects regions of moving objects and recover the depth map in a Markov Random Field framework. We first apply SIFT matching across frames in the video sequence and compute the camera parameters for all frames and the 3D positions of the SIFT feature points via structure from motion. Then, the 3D depths at these SIFT points are propagated to the whole image based on image over-segmentation to construct an initial depth map. Then the depth values for the segments with large re-projection errors are refined by minimizing the corresponding re-projection errors. In addition, we detect the area of moving objects from the remaining pixels with large re-projection errors. In the final step, we optimize the depth map estimation in a Markov random field framework. Some experimental results are shown to demonstrate improved depth estimation results of the proposed algorithm.

Keywords: Depth map recovery, structure from motion, Markov random field.

1 Introduction

In recent years, 3D imaging industry emerges rapidly. More and more movies take advantage of advanced 3D technology to reconstruct 3D motion or 3D scene for movie production or post-processing, which would produce amazing visual effect as if it were really happening. With the rapid development of the 3D movies, the 3D television also follows the trend. In each mall, we can see 3D TV exhibition in the spotlight. The digital camera also starts to include some 3D imaging function that supports 3D stereo images, 3D panorama shooting mode, and so on. In addition to these products, 3D digital photo frame and 3D digital printing are also the new related 3D products, and there will be more 3D products in the future.

For those 3D display monitors, it needs the image files that contain 3D information. If we can convert the captured 2D images or videos into the 3D format, we can view the images and videos with 3D display. Thus, the 2D-to-3D media conversion is very important for the development of the 3D imaging industry.

Most of the 2D-to-3D conversion technologies either cannot provide satisfactory results or require additional equipment, such as the stereo cameras or 3D range sensors. Although using the additional devices can provide more accurate results, they are more expensive and they cannot work for previous media. Therefore, we develop a new approach to estimating the depth map from a video sequence. The algorithm would automatically detect moving objects and estimate the depth map from a video sequence by considering both the spatial and temporal coherence in the video.

2 Previous Works

Depth estimation is an important and challenge problem in computer vision. The methods of depth map reconstruction can be roughly divided into two types: monocular vision based and multi-view based approaches. We briefly discuss them in this section.

2.1 Monocular Vision

In monocular vision, most methods used learning techniques or additional assumptions to obtain more information in a single image. Saxena et al. [1] proposed a learning algorithm to reconstructed 3D structure environment from a single image. They used supervised learning to learn how different visual cues are associated with different depths and used Markov Random Field (MRF) model to combine all the information. They also added other properties, including image features and a set of plane parameters of superpixels, into MRF to estimate the 3D positions and orientations for all superpixels. Unlike the previous approach which attempted to map from appearance features to depth, the algorithm in [2] first used semantic segmentation based on learning algorithm in an image to identify the positions of sky, tree, road, etc., in an image. They constructed a descriptor for each pixel, which includes local appearance features and global geometry features, and used a learning model to predict the initial depth map. Next they used some geometric constrains added to MRF framework to construct a smooth depth map. In addition, Hoiem [9] recovered the occlusion boundaries and depth ordering in the scene. They proposed a hierarchical segmentation process based on agglomerative merging to re-estimate boundary strength as the segmentation progresses.

2.2 Multi-view Vision

The most popular topic in multi-view vision is stereo reconstruction. Sun et al. [10] formulates the stereo matching problem in an MRF framework and solved it by belief propagation (BP) which works via iteratively passing messages to neighbor nodes. The drawback of this approach is that it spends too much time in processing the messages. Therefore, Felzenszwalb et al. [11] proposed a hierarchical BP algorithm to significantly reduce the computational complexity.

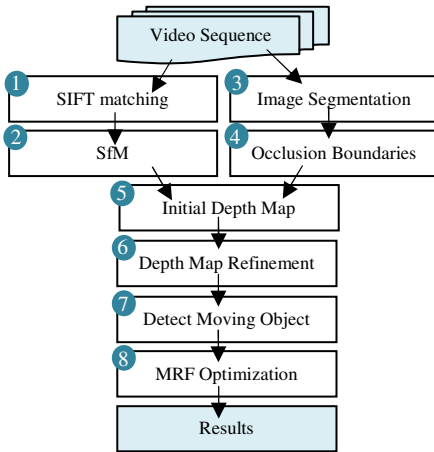


Fig. 1. The flowchart of the proposed method

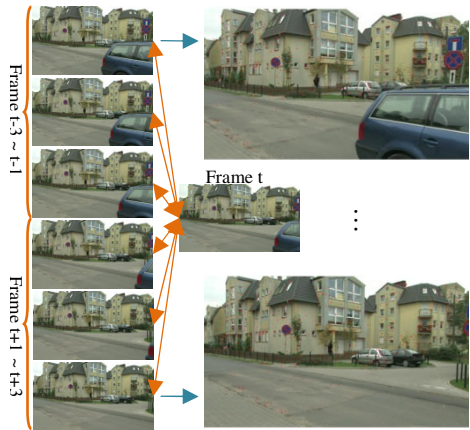


Fig. 2. The flowchart of SIFT matching

Another type of multi-view vision is for a video or multi-images captured from different views by using structure from motion (SfM) to compute the camera parameters for the multi-view images. Most of the SfM methods [5] defined energy functions on the 3D volume and used MRF to obtain the 3D surface. Recently, Zhang et al. [3, 4] used not only color constancy constraint but also the geometric coherence constraint which can maintain the temporal coherence in the video. For estimating the accurate disparity in textureless region, they added segmentation information based on color information to construct the initial disparity map. Next they used iteratively optimization which they called bundle optimization to refine the disparities in a pixelwise manner. More recently, Newcombe and Davison [6] used point-based structure from motion (SfM) to compute the camera poses first, computed the optical flow across frames, and used the triangulation method iteratively to optimize the 3D surface.

3 Proposed Method

Fig. 1 is the flow chart of the proposed depth map estimation algorithm. The goal for the first four steps is to collect information for recovering depth map; and the goal from 5th to 8th steps is to reconstruct and refine the depth map. First, we apply SIFT feature matching [12] to extract the corresponding points in the neighbor frames and use these points to compute the camera parameters by the SfM algorithm [13][14]. Then, we apply the mean-shift segmentation algorithm [7] and determine the occlusion boundaries based on [9] to classify the sky, ground, vertical regions. By using the 3D coordinate points from SfM, we construct a rough depth map based on the above over-segmentation. Next, we refine the depth map in the segments with large re-projection errors. However, the re-projection errors are still large after the depth refinement process in the regions containing moving objects. Thus, we use these regions as seeds to detect the regions of moving objects. In the final step, we optimize the depth map in the MRF framework to obtain accurate depth estimation.

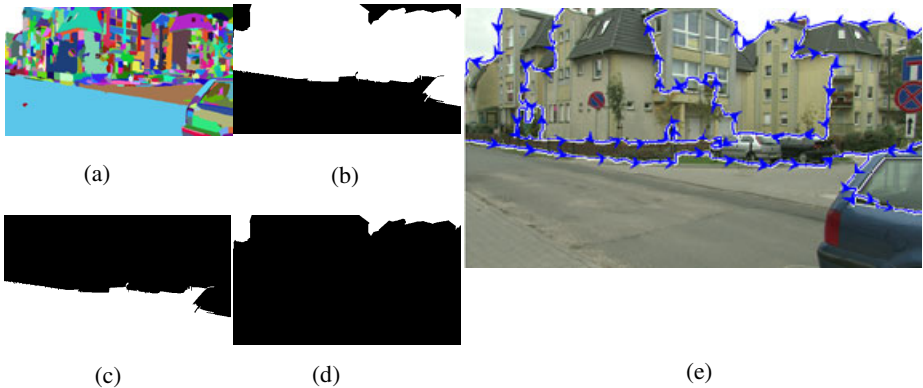


Fig. 3. (a) Mean-shift segmentation. (b-d) The mask of “vertical”, “ground” and “sky”. (e) Occlusion boundaries. The left side of the arrow is front of the right side of the arrow.

3.1 SIFT Matching

To construct the depth map from video, we would like to establish the relationship between the neighboring frames. The first step is to find reliable corresponding points in the video. Only having the corresponding points, we can calculate the camera parameters for all frames.

SIFT is an algorithm to detect and describe interest points in the images. The SIFT feature description is robust against the changes due to image scale, noise, illumination and rotation. For the above advantages, we apply the SIFT matching to compute the corresponding points across the six neighboring frames. Fig. 2 depicts the idea of the SIFT matching step.

3.2 Structure from Motion

In this step, we apply the SfM algorithm [13][14] to the SIFT correspondence points to estimate the camera parameters and 3D structures of these SIFT feature points. To estimate the depth map at a specific frame, I used seven neighboring consecutive frames into the SfM algorithm to compute the camera parameters and the 3D structures. These 3D points are the key to reconstruct the initial depth map in our algorithm. The details will be described in section 3.5.

3.3 Image Segmentation

Image segmentation is to divide an image into multiple homogeneous segments. The pixels in the same segment have similar visual characteristics, such as color, location, texture, etc. For these reasons, I use segmentation to help us improve the depth estimation, which will be described in section 3.6. In this paper, the mean-shift algorithm is chosen for the image over-segmentation, and an example result is depicted in Fig. 3(a). The segments are used for recovering the occlusion boundaries from an image, which will be discussed subsequently.

- 1. Set known 3D points**
 - 1.1 A sparse set of 3D points from SfM
- 2. Set depth to each segments**
 - 2.1 For each segment, construct a smooth depth map by (2) if it has enough information
 - 2.2 Propagate the depth value to neighbor segments
 - 2.3 Repeat step 2.1 and 2.2



Fig. 4. The flowchart of making initial depth map

Fig. 5. The initial depth map

3.4 Occlusion Boundaries

One of the important key in reconstructing depth map is occlusion relation. Owing to the change of view, some objects would be occluded by something in front of them, which would cause some problems in depth estimation. For example, sky is always covered by houses, trees, and so on. When we project a pixel to other frames, the labels may be different due to occlusion. This would lead to large errors in the image re-projection.

Hoiem et al. proposed an algorithm [9] to recover the occlusion boundaries and detecting the regions of “sky”, “vertical” and “ground” from an image. We incorporate their results of the occlusion boundaries and segment type classification into the depth map estimation framework to optimize the depth map estimation. An example of applying the algorithm in [9] is depicted in Fig. 3.

3.5 Initial Depth Map

We have collected a lot of useful information for reconstructing depth map in the previous steps. Therefore, we use all the information to construct an initial depth map. The main idea is that the neighboring pixels which have similar colors may have similar depth. In the SfM step, in addition to the camera parameters for all frames, we also obtain a sparse set of 3D feature points. Thus, these 3D points are used to propagate to the entire image. The flowchart of this step is shown in Fig. 4.

First, we scan all the 3D points to decide the range of the depth in the current frame, and record which pixels are already assigned with depth values. Then, for each segment, we counted how many pixels are already assigned with depth values. If the number is less than a threshold (5 in our implementation), we do nothing because there is not enough information to assign the depth for this segment. Otherwise, we consider each segment as a plane, and construct a smooth depth map for the segment. Hence, we set a linear plane, as in eq. (1), for each segment. Note that x and y are the coordinates of the pixel, a , b , c are the plane parameters, and d is the depth value. By using the least-square fitting, we calculate parameter a , b , c from some known points, $(x_1, y_1) \dots (x_n, y_n)$ with the relation given in eq. (2).

$$d = ax + by + c \quad (1)$$

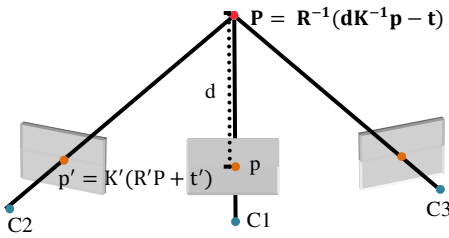


Fig. 6. The chart of projecting to the neighbor frame



Fig. 7. The depth map after refining

$$\begin{bmatrix} x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots \\ x_n & y_n & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix} \tag{2}$$

Not all segments have enough information to construct the depth. Hence, we propagate the depth from the known segments to unknown segments. In the edge of known segment and unknown segment, we check the similar degree of color. If the color is similar, then set the same depth value to the neighboring pixels. The steps of constructing depth and propagating depth are iterative. Fig. 5 depicts some results in this step. The red regions represent no depth information.

3.6 Depth Map Refinement

In the previous step, the initial depth map is rough and there is no depth information in some segments. Hence, in this step I would correct those segments which cause big error or have no depth information in the depth map.

First, I compute the projection error for each pixel in the whole image. The processing of projection is shown in Fig. 6. The blue points are the camera position and the orange points are the 2D coordinate in each frame. The red point is the corresponding 3D point of the orange points in the real world. Assuming the current frame is in the center of Fig. 6. p is one of pixel in the current frame. Giving the depth value of p and camera parameters of current frame and target frame, we can compute the corresponding coordinates, p' , in target frame by (3).

$$P = R^{-1}(dK^{-1}p - t) , p' = K'(R'P + t') \tag{3}$$

where p is the 2D image coordinate of a pixel, represented as $[x \ y \ 1]'$ in the homogeneous coordinate, P is the 3D coordinate of p in the world coordinate, represented as $[X \ Y \ Z]'$, K, R, t are the camera parameters at the current frame, K', R', t' are the camera parameters at the target frame, and p' is the corresponding coordinate of p in the target frame. If the depth value is accurate, then p and p' must have similar colors. Therefore, we define the re-projection error as a measure of difference between the colors at p and p' . Here, we use l_2 -norm to measure the difference between two colors in our implementation.

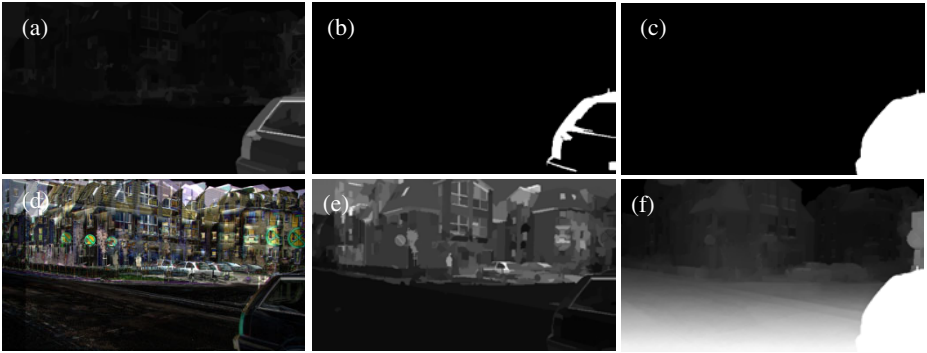


Fig. 8. (a) is the average error for each segments after depth refined. (b) is the seeded segments. (c) is the full region of moving object. (d) is the final error map based on pixels. (e) is the final error map based on segments. (f) The depth mp after optimization of MRF.

Due to the occlusion problem around the object boundary, we compute the average re-projection error for each segment instead of a pixel. If the average re-projection error in one segment is larger than a threshold, it denotes the depth value in this segment is inaccurate. For these segments, we search a new depth value within the depth range by finding the minimum re-projection error for the segment. Fig. 7 depicts an example of the depth map after the above refinement process.

3.7 Moving Object Detection

Most of the proposed approaches for the depth map reconstruction from a video sequences, such as [3], [4], and [6], do not account for moving objects in the video sequence. They all assume the scene is static; but in practice it is very common for the videos to contain moving objects.

It is not suitable to compute depth of moving objects by multi-view camera projections due to the unknown object motion. From the previous step, we have the average re-projection error for each segment after the depth map refinement, as shown in Fig. 8(a). However, there are still some segments with large errors, which means these segments are likely to be the moving object. Thus, we use these segments, as shown in Fig. 8(b), as the seeds to determine the regions of moving objects. Not all segments belonging to moving objects have large re-projection errors, just like the body of car in Fig. 8(a). Next, all the SIFT matching points in the seeded segments are used to calculate the displacement of moving object in the neighboring frames. The displacement is the difference of two corresponding points in x and y axes. To assure robust displacement estimation, we choose the median of all the displacements instead of mean. The relation equations are given in eq. (4) and (5), where p and p' are the corresponding points in the current frame and a neighboring frame, D_x and D_y are the displacements in x and y axes, and \widehat{D}_x and \widehat{D}_y are the final displacement.

$$D_{xi} = p_{xi} - p'_{xi} , \quad D_{yi} = p_{yi} - p'_{yi} \quad (4)$$

$$\widehat{D}_x = \text{median}_i (D_{xi}) , \widehat{D}_y = \text{median}_i (D_{yi}) \tag{5}$$

Because we assume the moving object in the neighboring frame and current frame would be the same shape and not be deformed. Then, we shift the neighboring frame by the displacement and subtraction by the current frame. We call the result as an error map. In our experiment, we use two neighboring frames, so we have two error maps. These two error maps are summed to form a combined error map. The final error map in Fig. 8(d) shows the probability of moving object. The pixels with darker color are more likely to be the moving object.

The process of finding the moving object is also based on segments. Therefore, the combined error map based on segments is depicted in Fig. 8(e). We use the seeded segments as the center, and visit the neighboring segments to merge the neighboring segment if the associated error is small. We repeat this segment merging process iteratively until convergence. Fig. 8(c) shows the region of detected moving object in this step.

3.8 MRF Optimization

The previous steps of reconstructing and refining depth map are all segment-based. Because these steps do not account for the relationship between neighboring segments, the computed depth map usually contains obvious errors. To improve the depth map estimation, we integrate the information in a Markov Random Field (MRF) framework to optimize the depth map.

Let the depth map of the current frame be represented as D . The image of the current frame is denoted by I , and $\widehat{I} = \{I'_t | t = 1, \dots, n\}$ is the set of reference frames. Then, we define the following energy function for MRF:

$$E(\widehat{D}; I, \widehat{I}) = E_D(\widehat{D}; I, \widehat{I}) + \alpha E_S(\widehat{D}; I) \tag{6}$$

where the data term E_D measures the projection error, the smoothness term E_S represents the smoothness on the depth map between neighboring pixels, And α is the weight used to balance these two terms. \widehat{D} is the refined depth map by MRF.

The definition of the data term $E_D(\widehat{D}; \widehat{I})$ is given in eq. (7). It computes the color difference between the corresponding points in the neighboring frames based on the current depth estimate.

$$E_D(\widehat{D}; I, \widehat{I}) = \sum_{p \in I} \sum_{t=1}^n \min(\text{abs}(I(p) - I'_t(p_t)), \delta) \tag{7}$$

$$p_t = K_t(R_t(R^{-1}(\widehat{D}(p))K^{-1}p - t) + t_t), \tag{8}$$

p_t is the corresponding point of a pixel p in the t^{th} reference frame, and δ is a threshold for the color difference. The symbols K_t, R_t, t_t denote the camera parameters at the t^{th} reference frame, and K, R, t are the camera parameters at the current frame. $\widehat{D}(p)$ is the depth value of p .

The smoothness term is defined by

$$E_s(\hat{D}; I) = \sum_{p_1 \in I} \sum_{p_2 \in N(p_1)} \frac{\text{abs}(\hat{D}(p_1) - \hat{D}(p_2))}{\text{abs}(I(p_1) - I(p_2)) + 1} \quad (9)$$

If p_1 and p_2 are neighbors, then their absolute difference in depth are divided by the corresponding absolute color difference, as in the denominator, and then added into smoothness term.

There are many algorithms based on MRF framework [16], and we choose graph-cut with swap algorithm [17][18] to refine our final depth map by minimizing the energy function given in eq. (6). The pixels belonging to moving objects, as determined in section 3.7, are not involved in the depth map refinement. After the optimization, we set an appropriate depth value to each of these regions of moving objects. We assume the moving object is usually vertical on the ground; therefore we assign the depth values of the pixels on the bottom of the moving object region as the depth for the region. An example of the MRF optimization result is depicted in Fig. 8(f).

4 Experiment Results

We apply in the proposed depth estimation algorithm to six real data sets. The data sets can be classified into two types: static scene and the scene containing moving objects. There are three videos belonging to static scene, i.e. “Road”, “Stair” and “Temple” [3][4], and we will compare the results with those in [3][4]. There are three videos containing moving objects. The videos “lovebird1” and “lovebird2” come from [19] and “Poznań” comes from [20].

The experimental results on the dataset “Poznań” are shown in Fig. 9-11. We focus on the passing vehicle in this video. In addition, there is a pedestrian in Fig. 9-11, but we only detect the moving person in Fig. 9. The key is the speed of the person is too slow to be detected as a moving object. Nevertheless, the result of depth map estimation in this example is acceptable. For the three video sequences, we detect the moving object successfully, and this helps to recover accurate depth maps.



Fig. 9. The 77th frame in the data set “Poznań”. (a) The reference frames. (b) The current frame. (c) The white region is the detected moving object. (d) Final depth map.

For the other datasets, we focus on the moving people in “lovebird1” and “lovebird2” videos. However, because of the high frame rate in these two videos, the movement of people is quite small in the neighboring frames. Therefore, we cannot detect moving object in these two videos, but it does not affect our results. The final estimated depth maps are depicted in Fig. 12. Although there is slight movement in these two videos, we still can obtain accurate depth map estimation.

Although we focus on handling moving object in this paper, we also test our algorithm on the three datasets of static scene. The results are shown in Fig. 13 and we also compare the depth estimation results with those by the state-of-the-art method [4]. From the figure, we can see our results are comparable to the results from [4] in these experiments for the static scene.



Fig. 10. The 130th frame in the data set “Poznań”. (a) The reference frames. (b) The current frame. (c) The white region is the detected moving object. (d) Final depth map.



Fig. 11. The 427th frame in the data set “Poznań”. (a) The reference frames. (b) The current frame. (c) The white region is the detected moving object. (d) Final depth map.

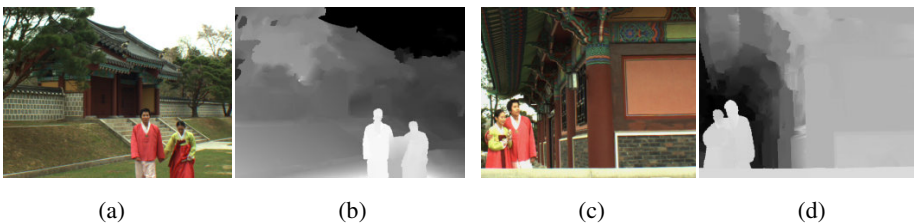


Fig. 12. The (a)&(c) images and (b)&(d) estimated depth maps for the 4th frame in the “lovebird1” and “lovebird2” sequences, respectively

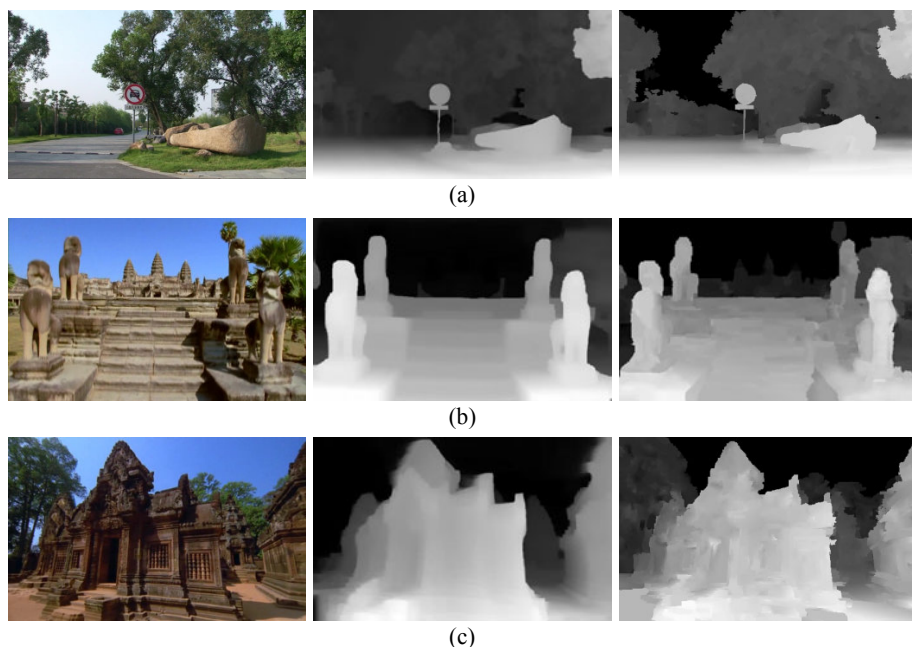


Fig. 13. From left to right: an image sample, depth map estimated by [4], depth map estimated by the proposed algorithm for the (a) Road, (b) Stair, and (c) Temple video sequences, respectively

5 Conclusion

In this paper, we proposed a robust system that reconstructs depth map from a video sequence automatically. In our system, we employ several computer vision technologies to help us construct accurate depth map from video. We integrate SIFT feature matching, SfM, mean-shift over-segmentation, occlusion boundary analysis to obtain some 3D information. To deal with moving object, we detect the segments of moving objects by selecting pixels with large re-projection errors as seeds in an iterative merging process. Finally, we integrate the 3D information in an MRF formulation to optimize the depth map. We demonstrate the proposed algorithm can provide satisfactory depth estimation results for videos with moving objects.

References

1. Saxena, A., Sun, M., Ng, A.Y.: Make3D: Learning 3D Scene Structure from a Single Still Image. *IEEE Trans. on Pattern Analysis and Machine Intelligence* (2008)
2. Liu, B., Gould, S., Koller, D.: Single Image Depth Estimation From Predicted Semantic Labels. In: *CVPR 2010* (2010)
3. Zhang, G., Jia, J., Wong, T., Bao, H.: Recovering Consistent Video Depth Maps via Bundle Optimization. In: *CVPR* (2008)

4. Zhang, G., Jia, J., Wong, T., Bao, H.: Consistent Depth Maps Recovery from a Video Sequence. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 31(6), 974–988 (2009)
5. Seitz, S.M., Curless, B., Diebel, J., Scharstein, D., Szeliski, R.: A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms. In: *CVPR* (2006)
6. Newcombe, R.A., Davison, A.J.: Live Dense Reconstruction with a Single Moving Camera. In: *CVPR* (2010)
7. Comanicu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence* (May 2002)
8. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient Graph-Based Image Segmentation. *International Journal of Computer Vision* 59(2) (September 2004)
9. Hoiem, D., Efros, A.A., Hebert, M.: Recovering Occlusion Boundaries from an Image. In: *IJCV* (2010)
10. Sun, J., Shum, H.Y., Zheng, N.N.: Stereo Matching Using Belief Propagation. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) *ECCV 2002*. LNCS, vol. 2351, pp. 510–524. Springer, Heidelberg (2002)
11. Felzenszwalb, P., Huttenlocher, D.: Efficient belief propagation for early vision. In: *IJCV*, pp. 1–8 (2007)
12. Pele, O., Werman, M.: A Linear Time Histogram Metric for Improved SIFT Matching. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part III*. LNCS, vol. 5304, pp. 495–508. Springer, Heidelberg (2008)
13. Martinec, D., Pajdla, T.: 3D Reconstruction by Fitting Low-Rank Matrices with Missing Data. In: *CVPR 2005*, pp. 198–205, IEEE (June 2005)
14. Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., Koch, R.: Visual modeling with a hand-held camera. *Intern. Journal of Computer Vision* 59(3), 207–232 (2004)
15. Alsabti, K., Ranka, S., Singh, V.: An Efficient k-means Clustering Algorithm. *Pattern Recognit. Lett.* 14(10), 763–769 (1993)
16. Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, A., Tappen, M., Rother, C.: A Comparative Study of Energy Minimization Methods for Markov Random Fields. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) *ECCV 2006, Part II*. LNCS, vol. 3952, pp. 16–29. Springer, Heidelberg (2006)
17. Boykov, Y., Veksler, O., Zabih, R.: Fast Approximate Energy Minimization via Graph Cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(11), 1222–1239 (2001)
18. Kolmogorov, V., Zabih, R.: What Energy Functions can be Minimized via Graph Cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26(2), 147–159 (2004)
19. Um, G., Bang, G., Hur, N., Kim, J., Ho, Y.-S.: Test Sequence “Lovebird1&2”
20. Domański, M., Grajek, T., Klimaszewski, K., Kurc, M., Stankiewicz, O., Stankowski, J., Wegner, K.: Poznań Multiview Video Test Sequences and Camera Parameters. *ISO/IEC JTC1/SC29/WG11 MPEG 2009/M17050*, Xian, China (October 2009)

An Iterative Algorithm for Efficient Adaptive GOP Size in Transform Domain Wyner-Ziv Video Coding

Khanh DinhQuoc, Xiem HoangVan, and Byeungwoo Jeon

School of Electrical and Computer Engineering, Sungkyunkwan University
300 Chunchun-dong, Jangan-gu, Suwon, 440-746, Korea
{diqkhanh,xiemhoang}@gmail.com, bjeon@skku.edu

Abstract. Transform Domain Wyner-Ziv Video Coding (TDWZ) is one of the most popular paradigms of Distributed Video Coding (DVC) which supports low encoding complexity. However, there is still a gap in its coding performance compared to conventional video coding standards such as MPEG-x, or H.264/AVC. In order for TDWZ to reach comparable performance to them, a good method for deciding a proper Group of Picture (GOP) size is in great necessity. From this point of view, we propose an iterative algorithm which efficiently determines GOP size based on an intra mode decision method at frame level. This approach firstly constructs a coarse GOP size and then refines it by iterative checking for final GOP size. Experimental results show superiority of the proposed algorithm with improvement up to 2dB in term of coding efficiency.

Keywords: Distributed Video Coding, Adaptive GOP size, Intra mode decision, Hierarchical structure.

1 Introduction

In human life, visual perception plays an essential role in receiving information from the outside world. That explains why video is the most attractive form of data in multimedia system. However, uncompressed video data require extremely large bandwidth for transmission or enormous storage capacity. To solve this problem, a lot of efforts have been made to develop video coding compression techniques which help reducing the number of bits to represent a video sequence.

Since the first digital video compression standard named H.120 was published by CCITT (Consultative Committee International Telephone and Telegraph) in 1984, the former name of ITU-T (International Telecommunication Union - Telecommunication Standardization Sector), video compression techniques have undergone long and steady developments. ISO/IEC (International Organization for Standardization/International Electrotechnical Commission) published MPEG-1, MPEG-2, and MPEG-4. ITU-T has also developed H.261, H.262, H.263 independently or jointly with ISO/IEC. Most recently, under the joint effort of both ISO/IEC and ITU-T, H.264/AVC (Advanced Video Coding) was finalized in 2003, and its a few extensions, namely, professional extension, scalable extension, multiview extension follow later. Currently, HEVC (High Efficiency Video Coding) is being developed under

JCT-VC (Joint Collaborative Team on Video Coding) with a target of compression efficiency enhancement by additional 50% compared to H.264/AVC inter coding in high profile.

However, those conventional video coding schemes require extreme amount of computational complexity at the encoder. Therefore, it is not suitable for up-link applications where low-encoding complexity is a major requirement.

With the recent explosion of hand-set devices, visual sensors and cameras, from wireless low power surveillance to mobile visual sensor network, or from video conference with mobile camera to multi-view video entertainment, a very low complexity at the encoder becomes a more essential feature due to their limited power supply and desire to have low complexity. In conventional video compression, encoder performs motion vector search, coding mode decision, rate-control, rate-distortion optimization, etc.; that means most of computing load is located at the encoder. But for the emerging applications, motion vector search, the most complexity component at encoder, is desired to be shifted to the decoder. Distributed Video Coding (DVC) scheme is one possibility to address this kind of requirement.

DVC is based on two key theorems by Slepian-Wolf [2] and Wyner-Ziv [3]. Slepian-Wolf theorem proved that two correlated sources can be encoded independently without any loss in coding efficiency if they are jointly decoded by exploiting source statistics. In 1976, Wyner-Ziv theorem further proved similar source coding method using side information for lossy compression. In 2002, two practical paradigms of DVC have been introduced by B. Girod et al. [4] and Ramchandram et al. [5]. The paradigm introduced in [4] was further developed into TDWZ (Transform Domain Wyner-Ziv Coding).

In TDWZ, Side Information (SI) generation and Channel Noise Modeling (CNM) are the two key functional components, which mostly decide its coding efficiency. Both of them depend on the distance between Wyner-Ziv (WZ) frames and key frames, that is, the GOP (Group of Picture) size. Pereira et al. [9] studied relationship between GOP size and performance of TDWZ. They showed that the rate-distortion or quality of decoded sequence relies on motion-level of sequences and different motion level should choose different GOP size. With high motion sequences such as Soccer or Coastguard, smaller GOP size gives better coding efficiency. Contrarily, for low motion sequence, larger GOP size gives better results. However when GOP size is too long, performance also decreases, for example with Hall monitor, best GOP is around four. Therefore, a proper GOP size is very important in term of coding efficiency.

Ascenso et al. [10] decided GOP size based on motion activity inside sequence. Their approach is based on four different matrices: Difference of Histogram (DH), Histogram of Difference (HD), Block Histogram Difference (BHD), and Block Variance Difference (BVD). They found that matrices DH and HD are powerful in working at frame level and detecting changes in global motion such as scene changes; matrices BHD and BVD are suitable for detecting high motion sequences, sequences with local motion in statistic background and group consecutive frames with same characteristic of motion into together. However, their method requires four complex matrices which require encoder to execute much calculation. The advantage of low complexity encoder is inflicted.

By other approach, Yaacoub et al. [11] determined a GOP size which has the minimum ratio PSNR/rate. That means they created models to estimate rate and PSNR values of both WZ and Intra frame, and set a loop for finding total PSNR and total rate to obtain the ratio of PSNR/rate. Their method requires heavy encoding complexity. Besides, using a fixed model for all sequences may not be totally accurate; for example, the result for sequence Salesman is slightly worse than fixed GOP size of 3.

In our proposed method, we determine GOP size in two steps - the first step finds a coarse GOP size by progressively looking for a frame which has higher probability of being intra-coded, then, the second step refines it based on iterative checking algorithm to decide the final GOP size. The intra mode decision at frame level compares temporal correlation and spatial correlation as described in [13]. Our simulation results show that it can adapt to various motion content of sequences and brings improvement up to 2dB in term of coding efficiency.

This paper is organized as follows. In Section 2, TDWZ with adaptive GOP size architecture is briefly introduced. Our proposed method is presented in Section 3. The proposed method is experimentally verified in Section 4; and finally, in Section 5, conclusions and some suggestions for future work are drawn.

2 Transform Domain Wyner-Ziv Coding with Adaptive GOP

Transform domain Wyner-Ziv video coding (TDWZ) is one representative DVC scheme [12] in which a video sequence is divided into key and WZ frames according to GOP size; the first and the last frames of a GOP are key frames, and all the other frames inside GOP are encoded as WZ frames. The key frames are encoded with low complexity compression techniques such as JPEG or H.264/AVC intra.

In the proposed TDWZ with adaptive GOP size (see Fig. 1), input sequence is firstly processed by GOP size controller to construct a GOP size. WZ frames are divided into blocks of 4x4 and discrete cosine transform is applied to each block to convert pixel data into frequency domain. Quantizer removes visual perception redundancy by applying large quantization step size at high frequency and small step size at low frequency.

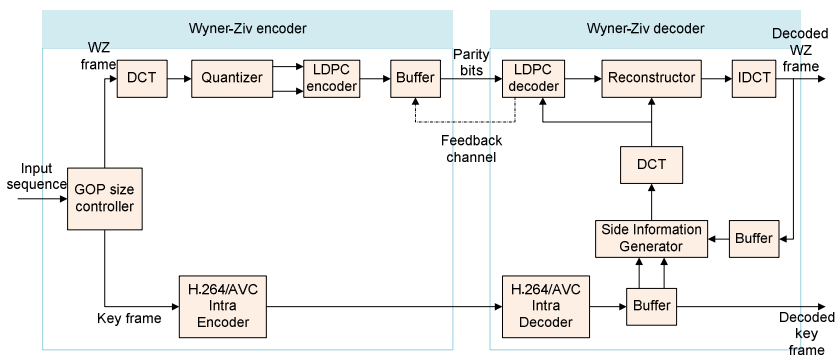


Fig. 1. Transform Domain Wyner-Ziv with GOP size controller

Then, DCT coefficients at the same frequency are grouped together and extracted into biplanes to feed into channel coding encoder.

The proposed TDWZ decoder decodes the first (\hat{X}_t) and the last frames ($\hat{X}_{t+GOPsize}$) of a GOP by H.264/AVC Intra decoder and store them in a buffer. Side information (SI) Generator refers to the decoded intra frames to create an error version ($Y_{t+WZindex}$) of WZ frame. Here, WZindex denotes a displaying order of a WZ frame. The procedure proposed by Ascenso et al. [6] is an efficient way to generate SI which uses bi-direction interpolation and weighted median filter. Based on backward reconstructed SI and forward reconstructed SI, a Channel Noise Model (CNM) between SI and WZ frames is made [14]. To correct error in SI, bitplanes of coefficients in frequency domain and channel model are processed at LDPC decoder to detect which bit has high probability of error. Then, requests for parity bits are sent to the encoder via a feedback channel as necessitated. Errors will be corrected by using transmitted parity bits. Under a hierarchical GOP structure, previously decoded WZ frame is used as a reference frame alternatively for a key frame which is located too far away from current WZ frame.

Although TDWZ is quite a powerful compression paradigm but it still cannot be comparable to the state-of-the-art performance of H.264/AVC. This gap must be overcome by researching more on coding efficiency improvement of TDWZ.

3 Proposed Iterative Algorithm for GOP Size Construction

The key idea of our proposed method is iteratively checking intra mode decision procedure. Firstly, we coarsely determine a GOP size as a preliminary one. Then by hierarchically checking frames inside the coarse GOP we figure out whether the preliminary GOP should be broken into smaller GOPs or not.

3.1 Intra Mode Decision

In DVC paradigm, quality of decoded WZ frame is still worse than that of decoded B frame in H.264. Sometimes it is worse than that of decoded intra frame, especially in high motion sequences such as Soccer or Stefan sequences. However, applying WZ coding is in general better in stationary frames. Therefore, there has been much effort to choose between intra coding and Wyner-Ziv coding to adapt to content of sequence [15-17]. Some researchers focused on using a threshold for mode decision, such as Belkoura et al. [15] or Do et al. [16]. However, using a same threshold, which highly depends on user's experience, for all sequences cannot give the best results always. Some others such as Ascenso et al.[17] tried to create models to estimate the rate and distortion for Wyner-Ziv mode and intra mode. Based on rate distortion model [17], they could judge which coding mode should be chosen. Nevertheless, such estimation model is not always correct for all sequences and may bring noticeable increment in encoding complexity. Xiem et al.[13] proposed a method of quite light complexity which did not rely on threshold. In this paper, we use similar idea as [13] as follows.

Denote $SAD_{t \rightarrow (t-i, t+j)}^T$ as a difference from the current frame at time t to two frames at time $t - i$ and $t + j$, respectively.

$$SAD_{t \rightarrow (t-i, t+j)}^T = \sum_{r=1}^H \sum_{c=1}^W \left\{ \left| F_{t-i}(r, c) - F_t(r, c) \right| + \left| F_t(r, c) - F_{t+j}(r, c) \right| \right\} \quad (1)$$

where $F_t(r, c)$ is a pixel value at position (r, c) of current frame t ; H and W respectively denote the height and the width of a frame. Positions of frames t , $t - i$ and $t + j$ are illustrated in Fig. 2. In general, increased motion level should lead to the difference increasing, and vice versa. Hence, $SAD_{t \rightarrow (t-i, t+j)}^T$ is an indicator (actually in a reverse sense) for temporal correlation. In the same spirit, compute SAD_t^S as a difference among blocks inside a frame at time t . Similarly, SAD_t^S is a representative of spatial correlation in a reverse sense.

$$SAD_t^S = \sum_{b=1}^B \sum_{r=1}^{blocksize} \sum_{c=1}^{blocksize} \left| F_{t,b}(r, c) - M_b \right| \quad (2)$$

where B denotes a total number of blocks in a frame t , $F_{t,b}(r, c)$ refers a pixel value at position (r, c) in the block b . In this paper, we used $blocksize=4$. M_b is a median value of three corresponding pixels of neighbor blocks: north (up), west(left) and northwest (top-left) direction. If SAD_t^S is larger than $SAD_{t \rightarrow (t-i, t+j)}^T$, frame t should be encoded as WZ frame and vice versa.

The proposed GOP size determination is made in two steps (coarse construction and iterative fine determination) as follows.

3.2 Coarse GOP Size Constructor

A key point of this step is that if SAD_t^S is the smallest than $SAD_{t \rightarrow (t-i, t+j)}^T$ values, the current frame should be coded as an intra frame. When the distances between current frame t to previous frame $t - i$ and next frame $t + j$ are larger, usually, the value of SAD_t^S is larger. We can infer that $SAD_{t \rightarrow (t-i, t+j)}^T > SAD_{t \rightarrow (t-1, t+1)}^T$ with every $i > 1$ and $j > 1$.

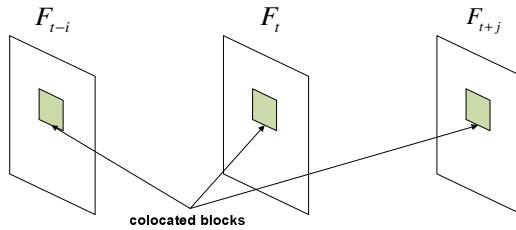


Fig. 2. Relative positions of frame t , $t - i$, and $t + j$

Therefore $SAD_{t \rightarrow (t-i, t+j)}^T$ corresponds to the largest difference around time t . As stated above, to identify a coarse GOP size, we search for a frame which has a smaller SAD_t^S than the smallest $SAD_{t \rightarrow (t-i, t+j)}^T$, and denote the distance between the current intra frame and the frame found as a coarse GOP size.

Besides, if GOP size is too long, the decoder must wait until finishing decoding all frames in this GOP, reorder frame index from coded index into display index, and then play entire GOP. Thus, a long GOP size take a long time to process and may undesirably affect easy video watching. In order to avoid such shortcomings, GOP size is better to be limited below some longest value, GOP_MAX. To identify the coarse GOP size, at first we compute $SAD_{t \rightarrow (t-i, t+j)}^T$ values of three consecutive frames and SAD_t^S values of center frame. Following, a coding mode is chosen according to the two relative factors explained above. If the center frame is decided to be intra-coded, this procedure is terminated and coarse GOP size is returned. Otherwise, we pick next three consecutive frames to check again in the same way. This procedure stops when center frame is assigned either as intra frame or GOP size reaches value of GOP_MAX. Fig. 4 summaries the coarse GOP size construction.

3.3 Iterative Finer-Size of GOP Determination Algorithm

When the previous key frame is too far away from the next key frame, $SAD_{t \rightarrow (t-i, t+j)}^T$ of current frame with respect to the previous key frame and the next key frame is larger. Thus, wrong motion vectors occur more frequently and degree of accuracy of bi-direction interpolation critically degrades. Consequently, quality of SI will be degraded then much more parity bits need to be sent to decoder to correct the worse SI. Therefore, overall TDWZ performance will be seriously affected. We can overcome this cascading problem by checking which mode is applied to the center frame by an intra mode decision procedure. In another word, we have two inequalities:

$$SAD_{t \rightarrow (previous_key, next_key)}^T > SAD_{t \rightarrow (t-1, t+1)}^T \quad \text{and} \quad SAD_{t \rightarrow (t-1, t+1)}^T < SAD_t^S$$

That means we cannot make a conclusion about relationship between $SAD_{t \rightarrow (previous_key, next_key)}^T$ and SAD_t^S . Therefore, for the next processed frame, we must check whether it should be encoded as intra mode or not by comparing the two SAD values at current frame. If $SAD_{t \rightarrow (previous_key, next_key)}^T > SAD_t^S$ the current frame is determined as an intra frame and we break the coarse GOP into two consecutive GOPs.

Besides, using a hierarchical GOP structure makes overall quality of decoded sequence better [8]. Inside an arbitrary GOP, the first decoded frame should be located in the center of GOP. Quality of decoded frame is increased by receiving parity bits and correcting errors in SI. Moreover, distance between the center frame and both key frames is smaller than distance apart themselves. By taking the analyzed advantages, we can improve the quality of latterly generated SIs which will increase overall GOP coding performance. We use this structure to order all frames fed into intra mode decision to guarantee that intra-coded frame (if it exists) will be as useful as possible.

Center frame of the coarse GOP in a hierarchical order is checked whether it should be intra-coded or WZ-coded. A coarse GOP is broken into two GOPs when center frame is decided as intra-coded frame; the iteration will continue applying into two halves of GOP with respecting them as two coarse GOPs. This process continues until all frames inside coarse GOP are assigned coding mode. For example, with a coarse GOP size of 8 as shown in Fig. 3 seven iterations are performed in order to mark from 1 to 7. Based on the assigned modes – Intra or WZ - for each frame in a coarse GOP, we separate it into smaller GOPs. Flow chart diagram of iterative finer-size algorithm is shown in Fig. 5.

In summary, the proposed iterative algorithm for adaptive GOP size works as follow. Procedure at section 3.2 finds a coarse GOP size. Final mode for a specific frame is determined by hierarchical checking of intra mode decision. At last, GOP size is determined by breaking the coarse GOP if necessary.

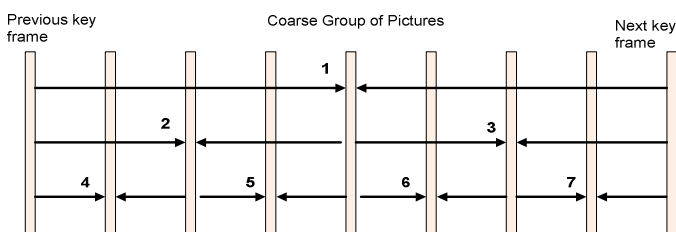


Fig. 3. Hierarchical structure for checking and encoding coarse GOP

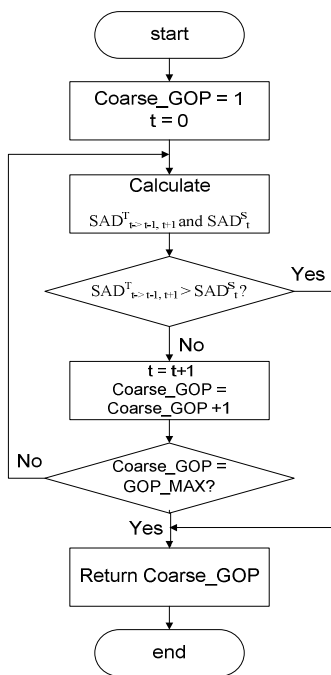


Fig. 4. Coarse GOP size determination

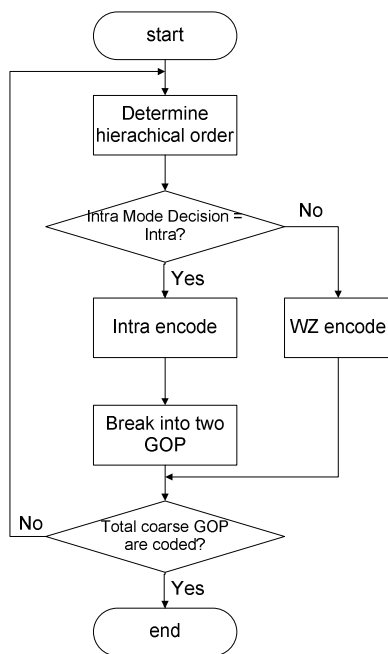


Fig. 5. Finer GOP determination

4 Experimental Results

Test condition

Our simulation is performed with three test sequences: Foreman, Hall monitor, and Soccer. The spatial and temporal resolutions are QCIF and 15 Hz, respectively. The number of frames is 149 frames for Foreman and Soccer, and 165 for Hall monitor. We compare coding performance of the proposed method ("**adaptive_GOP_size**") against following three methods.

- **SKKU DVC codec** which is developed by Digital Media Lab, Sungkyunkwan University [18]. GOP size of 2 is used for simulation.
- **DISCOVER codec** which was project result of EU IST FET program (Information Society Technologies – Future Emerging Technologies) [19]. It is chosen because it is the most well-known codec of Distributed Video Coding. GOP size of 2 is also used.
- **Intra Mode Decision** in [13] with GOP size of 2 and 4 is chosen to compare.

In addition, for displaying rate distortion performance, we set up quantization steps for Intra frames Q_p and quantization matrices for WZ frames Q_m . Our simulation is performed at four points of Q_p and Q_m pairs as shown in Table 1.

Table 1. Rate Distortion point for simulation

Foreman		Hall Monitor		Soccer	
Q_m	Q_p	Q_m	Q_p	Q_m	Q_p
1	40	1	40	1	40
4	34	4	34	4	34
7	29	7	29	7	29
8	25	8	25	8	25

Complexity

The proposed method ("**adaptive_GOP_size**") did not perform any motion search and compensation at the encoder to generate low quality Side Information. Also, it does not require any model for rate and distortion calculation. As described, our scheme assigned which frames should be intra-coded by just calculating difference from key frames and inner frames in GOP, and then based on it, an iterative algorithm fixes a final GOP size.

Consequently, there is not much calculation executed; thus encoding complexity is still kept very low. Fig. 6 illustrates that estimation of GOP size just occupies 4% in total encoding time. Such time usage for assigning GOP does not change much along with Q_m increase, although encoding time is a variant function of Q_m . This means that percentage of estimating GOP size even gets smaller when Q_m gets closer to its maximum value.

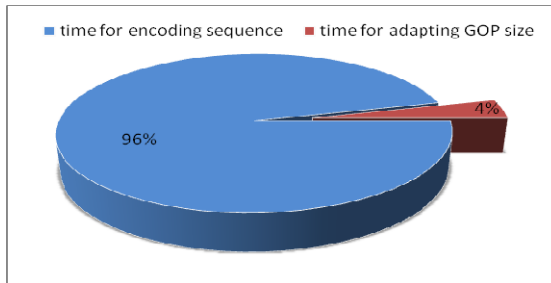


Fig. 6. Estimating GOP size time versus total encoding time (QM = 0, Hallmonitor)

Coding efficiency performance

The proposed method includes two steps to assign the final GOP size. For the sequence Foreman, GOP sizes after the first step – coarse GOP size construction – are shown in Fig. 7 (a). There is local motion at the first half of the sequence so coarse GOP sizes are around two or three. But around the 100th frame, high global motion occurs hence GOP size becomes equal to one. In rest part of the sequence which has very slow motion, longer GOP sizes are chosen. After that, these GOP sizes were refined and results were depicted in Fig. 7 (b) which makes one observe that GOPs are broken into two or more smaller GOPs (see the down arrows which point exactly to the frames changing from WZ frames into Intra frames.) After the second steps, GOP sizes are significantly changed.

For evaluating efficiency of the proposed method, we compared the rate distortion (RD) performance in Fig. 8, 9, and 10. It is clear that RD performance is improved remarkably with about over 2 dB increment at sequence Hallmonitor or Soccer, and about 0.9 dB increment at sequence Foreman compared to SKKU results without adaptive GOP size [18]. However, when motion level increases, the gap between the simulation results and [13] decreases. In case of low motion level sequence, ratio of WZ frames should increase but it is limited because of fixed GOP size. This situation changes when motion level raises, most odd frames are encoded as Intra frames and it approaches the best GOP size. Indeed, our proposed method has shown that, the best

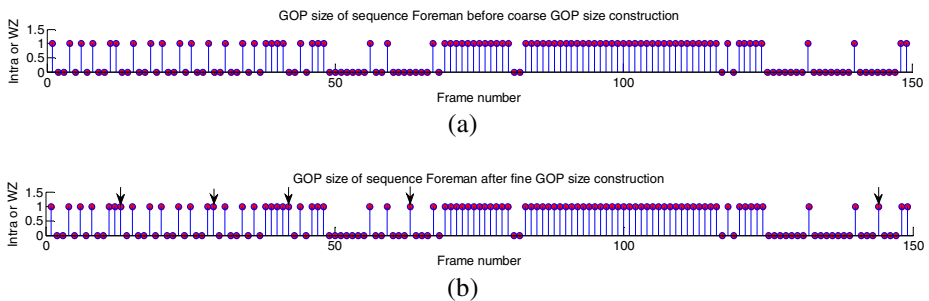


Fig. 7. Effect of the proposed iterative fine-size determination algorithm

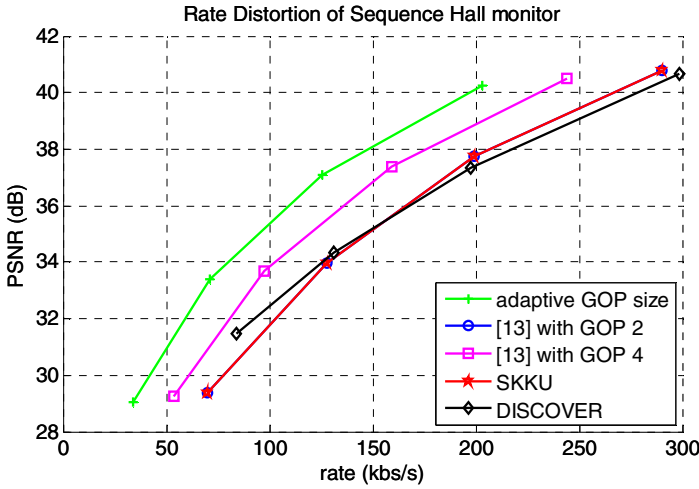


Fig. 8. Rate distortion performance of sequence Hall monitor

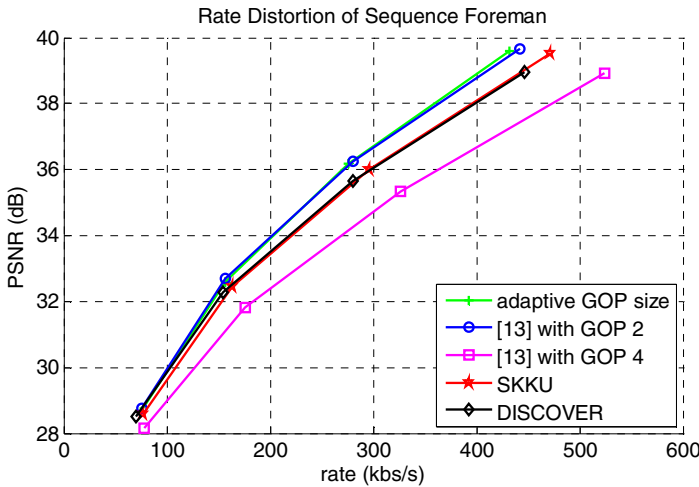


Fig. 9. Rate distortion performance of sequence Foreman

mode for sequence Soccer is almost Intra encoded. Simulation results also proved the stated analysis, and our proposed method is better than [13] by amounting to 2.2 dB in case of GOP equal to 2, and 1.2 dB with GOP equal to 4 when we examine the sequence Hallmonitor. In case of the sequence Soccer, the increment is about 0.2 dB against [13].

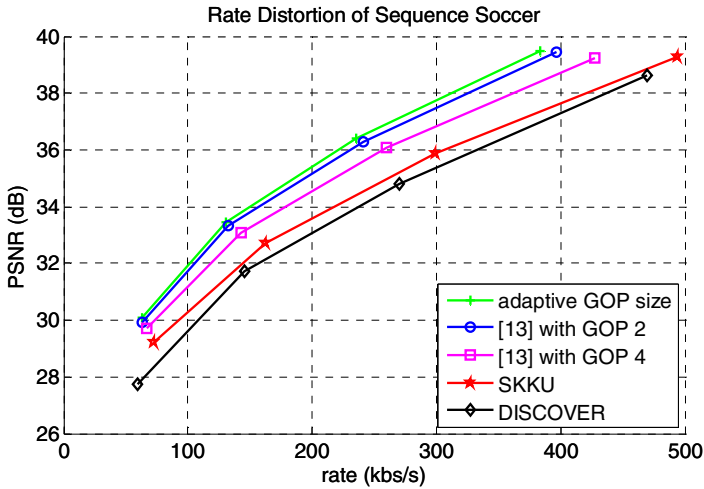


Fig. 10. Rate distortion performance of sequence Soccer

5 Conclusions and Future Works

This paper proposes an iterative algorithm for constructing GOP size in transform domain Wyner-Ziv video coding. This novel method is developed based on intra frame mode selection at frame level and iterative technique. Hierarchical structure is used for better results. Experimental results showed the superiority of the proposed method compared to previous works both in coding efficiency with improvement up to 2dB and low encoding complexity.

Because the simulation results still cannot reach the performance of H.264/AVC, there are much work remained to be done to reduce this gap. In the future works, an improved method for intra frame mode selection will be investigated to increase the accuracy of GOP size construction.

Acknowledgments. This work was supported by the IT R&D program of MKE/KEIT. [KI002142, Development of Ultra Low Complexity Video Coding Technique for Next-generation Mobile Video Service.

References

1. Wiegand, T., Sullivan, G.J., Bjontegaard, G., Luthra, A.: Overview of the H.264/AVC video coding standard. *IEEE Trans. on Circuits and Systems for Video Technology* 13, 560–576 (2003)
2. Slepian, D., Wolf, J.K.: Noiseless coding of correlated information sources. *IEEE Trans. on Inform. Theory* IT-19, 471–480 (1973)
3. Wyner, D., Ziv, J.: The rate-distortion function for source coding with side information at the decoder. *IEEE Trans. on Inform. Theory* IT-22, 1–10 (1976)

4. Aaron, A., Zhang, R., Girod, B.: Wyner-Ziv Coding of Motion Video. In: Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA (2002)
5. Puri, R., Ramchandran, K.: PRISM: A New Robust Video Coding Architecture Based on Distributed Compression Principles. In: 40th Allerton Conference on Communication, Control and Computing, Allerton, USA (2002)
6. Ascenso, J., Brites, C., Pereira, F.: Motion Compensated Refinement for Low Complexity Pixel Based Distributed Video Coding. In: 4th Conference on Advanced Video and Signal Based Surveillance AVSS, Italy, (2005)
7. Brites, C., Ascenso, J., Pereira, F.: Studying Temporal Correlation Noise Modeling for Pixel Based Wyner-Ziv Video Coding. In: International Conference on Image Processing ICIP, USA (2006)
8. Ascenso, J., Pereira, F.: Hierarchical motion estimation for side information creation in Wyner-Ziv video coding. In: Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication (2008)
9. Pereira, F., Ascenso, J., Brites, C.: Studying the GOP Size Impact on the Performance of a Feedback Channel-Based Wyner-Ziv Video Codec. In: Mery, D., Rueda, L. (eds.) PSIVT 2007. LNCS, vol. 4872, pp. 801–815. Springer, Heidelberg (2007)
10. Ascenso, J., Brites, C., Pereira, F.: Content Adaptive Wyner-ZIV Video Coding Driven by Motion Activity. In: IEEE International Conference on Image Processing, Atlanta, GA (2006)
11. Yaacoub, C., Farah, J., Pesquet-Popescu, B.: New Adaptive Algorithms for GOP Size Control with Return Channel Suppression in Wyner-Ziv Video Coding. In: 16th IEEE International Conference on Image Processing (2009)
12. Aaron, A., Rane, S., Setton, E., Girod, B.: Transform-domain Wyner-Ziv Codec for Video. In: Proc. SPIE Visual Communications and Image Processing, San Jose, USA (2004)
13. Xiem, H.V., Park, J., Jeon, B.: Flexible Complexity Control based on Intra Frame Mode Decision in Distributed Video Coding. In: Proc. of IEEE Broadband Multimedia Systems and Broadcasting, Germany (2011)
14. Brites, C., Pereira, F.: Correlation Noise Modeling for Efficient Pixel and Transform Domain Wyner-Ziv Video Coding. IEEE Trans. on Circuits and Systems for Video Technology 18, 1177–1190 (2008)
15. Belkoura, Z., Sikora, T.: Towards rate-decoder complexity optimisation in turbo-coder based distributed video coding. In: Picture Coding Symposium, Beijing, China (2006)
16. Do, T., Shim, H.J., Jeon, B.: Motion linearity based skip decision for Wyner-Ziv coding. In: 2nd IEEE International Conference on Computer Science and Information Technology, China, (2009)
17. Ascenso, J., Pereira, F.: Low complexity intra mode selection for efficient distributed video coding. In: IEEE International Conference on Multimedia and Expo, USA (2009)
18. Park, J., Jeon, B., Wang, D., Vincent, A.: Wyner-Ziv video coding with region adaptive quantization and progressive channel noise modeling. In: Proc. of IEEE Broadband Multimedia Systems and Broadcasting, pp. 1–6 (2009)
19. DISCOVER codec,
http://www.img.lx.it.pt/~discover/rd_qcif_15_gop2.html

A Robust Zero-Watermark Copyright Protection Scheme Based on DWT and Image Normalization

Mahsa Shakeri and Mansour Jamzad

Computer Engineering Department,
Sharif University of Technology, Tehran, Iran
mshakeri@ce.sharif.edu
Jamzad@sharif.edu

Abstract. Recently, protecting the copyright of digital media has become an imperative issue due to the growing illegal reproduction and modification of digital media. A large number of digital watermarking algorithms have been proposed to protect the integrity and copyright of images. Traditional watermarking schemes protect image copyright by embedding a watermark in the spatial or frequency domain of an image. However, these methods degrade the quality of the original image in some extent. In recent years, a new approach called zero-watermarking algorithms is introduced. In these methods, the watermark does not require to be embedded into the protected image but is used to generate a verification map which is registered to a trusted authority for further protection. In this paper a robust copyright proving scheme based on discrete wavelet transform is proposed. It uses a normalization procedure to provide robustness against geometric distortions and a cellular automaton for noise robustness. Experimental results on images with different complexity demonstrate that our proposed scheme is robust against common geometric and non geometric attacks including blurring, JPEG compression, noise addition, sharpening, scaling, rotation, and cropping. In addition, our experimental results obtained on images with different complexities showed that our method could outperform the related methods in most cases.

Keywords: Zero-watermarking, copyright protection, discrete wavelet transform, image normalization, cellular automata, robustness.

1 Introduction

With the rapid expansion of Multimedia and the networking technology, the problem of illegal reproduction and modification of digital media has become increasingly important. The watermarking technique is one of the solutions for image copyright protection. Traditional watermarking methods [1, 2, 3], embed a watermark logo in the spatial or frequency domain of an original image so that the watermark information can be extracted for copyright protection. However, this embedding procedure degrades the original image's quality. Therefore, a new watermarking approach called zero-watermarking has been proposed in which the quality of the original image does not change, but the watermark is constructed by extracting features of the original

image. Recently, a group of zero-watermarking methods have been introduced [4, 5, 6, 8, 9] that extract some binary features from the host image. Then by applying exclusive-or between these extracted features and the binary logo, a verification map is achieved. This verification map is registered in a trusted authority for ownership protection. In this paper, we propose a novel scheme from this group of zero-watermarking methods.

In [4], Chen et al. proposed a wavelet-based copyright proving scheme in which the features of host image were extracted from t-level LL subband. This scheme is not robust enough under some geometric attacks. A new wavelet-based zero-watermarking algorithm was proposed in [5] that achieved better robustness than [4] by applying a kind of smoothing filter on t-level LL subband. An adaptive approach proposed in [6] which allowed image owners to adjust the strength of watermarks through a threshold by using Sobel technology [7]. In [8] genetic algorithm is used to provide a proper threshold. As these two schemes depend on the edge features that are extracted from an image, the wrong threshold can increase false positive significantly. A copyright protection method for digital image with 1/T rate forward error correction was proposed in [9]. The watermark logo is fused with noise bits to improve the security and later became exclusive-or with the feature value of the image by 1/T rate forward error correction. Although this method has high performance on non geometric attacks, it still does not achieve high performance for geometric attacks such as rotation and shearing. In this paper, we use discrete wavelet transform and a normalization procedure to extract binary features and compare our scheme with [4], [5] and [9] schemes.

The rest of this paper is organized as follows: In section 2 discrete wavelet transform, cellular automata and the procedure of image normalization is presented. The proposed method is elaborated in section 3 followed by the simulation results and performance analysis in section 4. Finally, the conclusions are summed up in section 5.

2 Related Knowledge

2.1 Discrete Wavelet Transformation (DWT)

The discrete wavelet transformation is a mathematical tool for decomposing functions which can examine an image in the time and frequency domains, simultaneously. The transformed image is obtained by repeatedly filtering the image on a row-by-row and column-by-column basis. After each level wavelet transformation, the low frequency subband which contains most of the energy in the image may be transformed again. If the process is repeated t times, it is called t-level wavelet transformation and the low frequency subband is referred by LL_t . As the low frequency components in DWT are less sensitive under signal processing attacks than the high frequency components, we used the subband LL_t of the original image in our proposed method.

2.2 Cellular Automaton

A bi-dimensional cellular automaton is a triple $A = (S, N, \delta)$; where S is a nonempty set, called the state set, $N \subseteq \mathbb{Z}^2$ is the neighborhood, and $\delta : S^N \rightarrow S$ is the local tran-

sition function (rule); the argument of δ indicates the states of the neighborhood cells at a given time, while its value represents the central cell state at the next time.

A digital image is a bi-dimensional array of $n \times n$ pixels. Therefore, in this paper our model is based on a bi-dimensional symmetric cellular automata of the form $A = (S, N \delta)$ with $S = \{\#, 0, 1, \dots, k - 1\}$ for an image with k gray levels. $\#$ is the quiescent state associated to the cells outside the grid. N is the Moor neighborhood (the eight neighboring cells surrounding a cell) while the transition function $\delta : S^9 \rightarrow S$ is defined as follows:

$$\delta((s_i)_{i=1}^9) = \begin{cases} \sum_{i=1}^9 s_i / 9, & \text{if } s_5 \neq \# \\ \#, & \text{if } s_5 = \# \end{cases} \tag{1}$$

A cell (not being in the quiescent state) changes its state to the mean state of the cells in the neighborhood. The cells disposed outside the lattice of $n \times n$ pixels are assumed to be in the quiescent.[10]

2.3 Affine Transform

An image $g(x,y)$ is said to be an affine transform of $f(x,y)$ if there is a matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ and vector $d = \begin{pmatrix} d_1 \\ d_2 \end{pmatrix}$ such that $g(x, y) = f(x_a, y_a)$, where

$$\begin{pmatrix} x_a \\ y_a \end{pmatrix} = A \cdot \begin{pmatrix} x \\ y \end{pmatrix} - d. \tag{2}$$

It can be seen that geometric transformation such as: rotation, scaling, translation, shearing are all the special cases of affine transform.

Lemma 1: If $g(x,y)$ is an affine transformed image of $f(x,y)$ obtained with affine matrix $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ and $d=0$, then the following identities hold:

$$m'_{pq} = \sum_{i=0}^p \sum_{j=0}^q \binom{p}{i} \binom{q}{j} a_{11}^i \cdot a_{12}^{p-i} \cdot a_{21}^j \cdot a_{22}^{q-j} \cdot m_{i+j,p+q-i-j} \tag{3}$$

$$\mu'_{pq} = \sum_{i=0}^p \sum_{j=0}^q \binom{p}{i} \binom{q}{j} a_{11}^i \cdot a_{12}^{p-i} \cdot a_{21}^j \cdot a_{22}^{q-j} \cdot \mu_{i+j,p+q-i-j} \tag{4}$$

where m'_{pq}, μ'_{pq} are the moments of $g(x,y)$, and m_{pq}, μ_{pq} are the moments of $f(x,y)$.

2.4 Image Normalization

The key idea of the proposed watermarking scheme is to extract binary features from a normalized image in both watermark embedding and extraction phases. For image normalization, five different affine transforms are applied to the original image $f(x,y)$ by which the geometric distortions of the image are eliminated [11, 12]. The normalization procedure for a given image $f(x,y)$ consists of five steps. Let m_{pq} and $\mu_{pq}, p, q = 0,1,2, \dots$ be the geometric and central moments of image, respectively. The normalization steps are as follows:

Step 1: The translation is eliminated by setting the center of the image $f(x, y)$ at point $\begin{pmatrix} d_1 \\ d_2 \end{pmatrix}$ with $d_1 = \frac{m_{10}}{m_{00}}$ and $d_2 = \frac{m_{01}}{m_{00}}$ in (2). m_{00} , m_{01} and m_{10} are the moments of $f(x, y)$. The resulting image is denoted by $f_1(x, y)$ [11].

Step 2: Apply a shearing transform to $f_1(x, y)$ in the x direction with matrix $A = \begin{pmatrix} 1 & \beta \\ 0 & 1 \end{pmatrix}$ and $d=0$ in (2). The parameter β is set in a way that the resulting image, denoted by $f_2(x, y)$ achieves $\mu_{30}^{(2)} = 0$ [11], where the superscript is used to denote $f_2(x, y)$. According to equation (4), we have

$$\mu_{30}^{(2)} = \mu_{30}^{(1)} + 3\beta\mu_{21}^{(1)} + 3\beta^2\mu_{12}^{(1)} + \beta^3\mu_{03}^{(1)}. \tag{5}$$

where $\mu_{pq}^{(1)}$ are the central moments of $f_1(x, y)$.

Step 3: Apply a shearing transform to $f_2(x, y)$ in the y direction with matrix $A = \begin{pmatrix} 1 & 0 \\ \gamma & 1 \end{pmatrix}$ and $d=0$. The parameter γ is set in a way that the resulting image, denoted by $f_3(x, y)$ achieves $\mu_{11}^{(3)} = 0$ [11]. From identity (4), we have

$$\mu_{11}^{(3)} = \gamma\mu_{20}^{(2)} + \mu_{11}^{(2)}. \tag{6}$$

Step 4: Scale $f_3(x, y)$ in both x and y directions with $A = \begin{pmatrix} \alpha & 0 \\ 0 & \delta \end{pmatrix}$ and $d=0$. The parameters α and δ are determined in a way that the resulting image denoted by $f_4(x, y)$ achieves a determined standard size and its moments become $\mu_{50}^{(4)} > 0$ and $\mu_{05}^{(4)} > 0$ [11].

Step 5: Apply a rotation transform to $f_4(x, y)$ with matrix $A = \begin{pmatrix} \cos\emptyset & \sin\emptyset \\ -\sin\emptyset & \cos\emptyset \end{pmatrix}$ and $d=0$. The parameter \emptyset is determined in a way that the resulting image denoted by $f_5(x, y)$ achieves $\mu_{03}^{(5)} + \mu_{21}^{(5)} < 0$ [12]. The image $f_5(x, y)$ is the normalized image that is used in our watermarking method.

Theorem 1: An image $f(x, y)$ and its affine transform have the same normalized image.

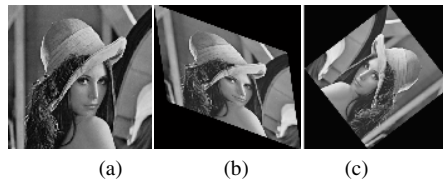


Fig. 1. (a) Original image Lena. (b) The Lena image after shearing distortion in x and y directions. (c) Normalized image from both (a) and (b).

This theorem has been proved in [11]. Fig. 1 illustrates the normalization procedure. The original image “Lena” is shown in Fig. 1(a). Also, this image after affine distortion is shown in Fig. 1(b). Both of these images yield the same normalized image, illustrated in Fig. 1 (c), when the above normalization procedure is applied.

In our proposed scheme, we first obtain the normalized image and then extract binary features so that our method becomes robust against geometric distortions because according to *Theorem 1* and as shown in Fig. 1, an image and its affine transform (image with geometric distortions) have the same normalized image.

3 The Proposed Method

A zero-watermarking method consists of two phases: the signature extraction procedure and the watermark verification procedure. In the first phase some binary features are extracted from the original image. Then bitwise exclusive-or is applied on the binary map image and a permuted binary logo to generate the verification map. Finally, the verification map that is a kind of signature is sent to a trusted authority. The second phase consists of retrieving the binary logo and authenticating the copyright of the image. Fig. 2 shows the steps of the signature extraction procedure and Fig. 3 illustrates the verification procedure. The detailed description of these two phases is stated in sections 3.1 and 3.2.

3.1 Signature Extraction Procedure

Let the original image O be a gray level image of size $H_O \times W_O$ and the digital logo W be a binary image of size $H_w \times W_w$. The original image O and the binary logo W are defined as follows:

$$O = \{o_{ij} | 0 \leq o_{ij} \leq 255, 0 \leq i \leq H_O, 0 \leq j \leq W_O\}. \tag{7}$$

$$W = \{w_{ij}, w_{ij} \in \{0,1\}, 0 \leq i \leq H_w, 0 \leq j \leq W_w\}. \tag{8}$$

Step 1) *Choosing the appropriate level of wavelet transform*: Assuming the original image and the binary logo are of size 2^n , the original image will be decomposed by using t level wavelet transformation in step 3 to obtain the low frequency subband LL_t . The value of t is determined by the following formula:

$$2^t = \frac{W_O}{W_w} = \frac{H_O}{H_w}. \tag{9}$$

Step 2) *Normalization*: The normalization procedure which has been described in section 2.4 is applied on the original image to obtain the normalized image.

Step 3) *Discrete wavelet transformation of the normalized image*: The normalized image is decomposed by using t level wavelet transformation to obtain the low frequency subband LL_t . Let call it L and define it as follows:

$$L = \{L_{ij} | 0 \leq L_{ij} \leq 255, 0 \leq i \leq H_L, 0 \leq j \leq W_L\}. \tag{10}$$

where W_L and H_L are the width and height of L .

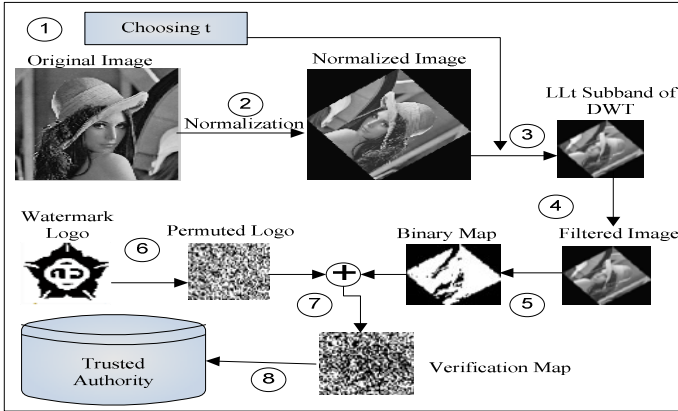


Fig. 2. The digital signature extraction procedure of our proposed method

Step 4) *Cellular Automata on L*: A two-dimensional cellular automata is used as described in section 2.2 for noise filtering over each pixel of the image L. The resulted image is called LS.

Step 5) *Binary map generation*: Let the gray level *avg* be taken as the average value of the image LS. Then the binary map *p* is generated as follows:

$$p_{ij} = \begin{cases} 0, & \text{if } LS(i, j) < avg \\ 1, & \text{if } LS(i, j) \geq avg \end{cases} \quad (11)$$

Step 6) *Logo permutation*: To increase randomness, the logo *W* is scrambled by a two dimensional pseudorandom permutation with seed *s* [13]. The permuted logo is denoted by \hat{W} .

$$\hat{W} = \{\hat{w}_{ij} | \hat{w}_{ij} = PRP_s(w_{ij}), 0 \leq i, i \leq H_w, 0 \leq j, j \leq W_w\} \quad (12)$$

where PRP_s is the permuted function with seed *s*.

Step 7) *Verification map generation*: We apply bitwise exclusive-or between the binary map *p* and the permuted logo \hat{W} to generate the verification map *K* as follows:

$$K = P \oplus \hat{W}. \quad (13)$$

Step 8) *Digital signature and timestamping*: The resulted verification map *K*, the width and height of the original image, the seed *s*, the value *t* and the signer’s identity ID_{signer} is sent to a trusted authority. The trusted authority center generates the certificate C_{TA} for the host image as follows:

$$C_{TA} = Hash_{TA}(K || W_O || H_O || ID_{signer} || s || t || ts). \quad (14)$$

where $Hash_{TA}(\cdot)$ is the one-way hash function, *ts* is a timestamp and $||$ is the concatenation operator. At last, the trusted authority center publishes the certificate C_{TA} , timestamp *ts* and the hash function $Hash_{TA}(\cdot)$ on its bulletin board.

3.2 Watermark Verification Procedure

Step 1) *Getting the information from trusted authority*: First, the verifier receives the C_{TA} , timestamp ts and hash function $Hash_{TA}(\cdot)$ from trusted authority. Then C_{TA}^* is computed by using $\{K, W_O, H_O, ID_{signer}, s, t\}$ and ts as follows:

$$C_{TA}^* = Hash_{TA}(K || W_O || H_O || ID_{signer} || s || t || ts) . \tag{15}$$

If the resulted C_{TA}^* will not be the same as C_{TA} the verification procedure is terminated in this phase; otherwise, the validity of the information is proven.

Step 2) *Normalization*: Normalize the test image by applying the indicated normalization procedure.

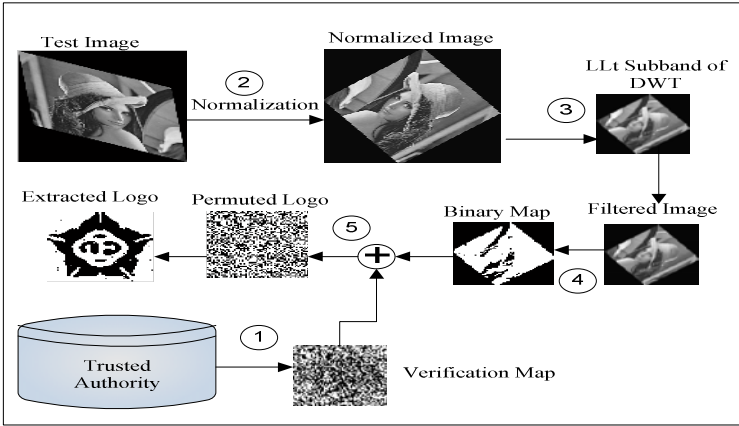


Fig. 3. Watermark verification procedure of our proposed method

Step 3) *Discrete wavelet transformation and filtering by Cellular automata*: The normalized test image is decomposed by performing t-level wavelet transform to obtain the subband LL_t . Then the cellular automata as described in section 2.2 is applied to achieve $\hat{L}S$.

Step 4) *Binary map generation*: The binary map \hat{p} is generated by calculating the average value of $\hat{L}S$ as noted in signing procedure.

Step 5) *Retrieving the logo*: By applying exclusive-or between the binary map \hat{p} and the verification map K the permuted logo \hat{W} is obtained as follows:

$$\hat{W} = \hat{p} \oplus K . \tag{16}$$

The inverse pseudorandom permutation of the permuted logo \hat{W} is computed by function PRP_s^{-1} and seed s . the retrieved logo \hat{W} is as follows:

$$\tilde{W} = \{\tilde{w}_{ij} | \tilde{w}_{ij} = PRP_s^{-1}(\tilde{w}'_{ij}), 0 \leq i, i \leq H_w, 0 \leq j, j \leq W_w\}. \tag{17}$$

Finally, the verifier can visually verify the accuracy of retrieved logo and validate the ownership of the test image.

4 Experimental Results and Analysis

We conducted some experiments to demonstrate the robustness of our copyright proving scheme against signal processing and geometric attacks. In addition, we compared our method with other related works.

4.1 Experimental Results

We measure the quality between the original image and the attacked image by using the peak signal to noise ratio (PSNR) which is defined as follows:

$$PSNR = 10 \log_{10} \frac{\max(O_{ij})^2 \times W_O \times H_O}{\Sigma(O_{ij} - \hat{O}_{ij})^2}. \tag{18}$$

where O and \hat{O} are the original image and the attacked image, respectively. W_O is the width and H_O is the height of image O .

After extracting the watermark the normalized cross correlation (NC) between the original watermark and the extracted logo is computed to evaluate the correctness of the extracted logo. The NC formula for binary watermark is defined as follows:

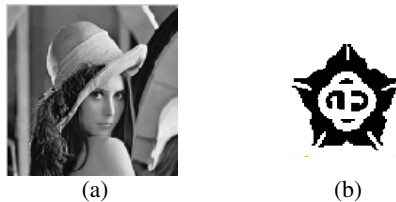


Fig. 4. (a) Test image 'Lena'. (b) Binary watermark logo.

$$NC(w, \hat{w}) = \frac{\Sigma_{i=1}^{H_w} \Sigma_{j=1}^{W_w} w_{ij} \times \hat{w}_{ij}}{\Sigma_{i=1}^{H_w} \Sigma_{j=1}^{W_w} w_{ij}^2}. \tag{19}$$

w_{ij} and \hat{w}_{ij} are the values that are located in the coordinate (i,j) of the original watermark w and the extracted watermark logo \hat{w} . Here w_{ij} is set to -1 if it is equal to 0. Otherwise, it is set to 1. The bits of \hat{w}_{ij} is set in the same way. Therefore, the value of NC will be in the range of [-1 1].

4.2 Applying Attacks to the Original Image

In this paper we used the same original image and binary logo as [4] and [5] for better comparison. Fig. 4 shows the original 256 gray-level Lena image of size 512×512 and the binary logo of size 64×64 .

Attack in zero-watermarking methods is defined as any kind of distortions that can be applied on the original image in order to disturb the procedure of binary feature extraction in watermark verification phase. Different signal processing and geometric attacks were applied on the test image and the Normalized cross correlation between the retrieved logo and the original logo is computed. In addition the related methods [4], [5] and [9] have been implemented and the results of our proposed method are compared with them. Following is the list of attacks that were applied to the test image.

- Attack 1.* Gaussian blurring with two pixels radius.
- Attack 2.* JPEG compression with compression rate 13% .
- Attack 3.* Sharpening the test image.
- Attack 4.* 7% Gaussian noise .
- Attack 5.* Reducing the size to 128×128 then resizing it to 512×512
- Attack 6.* 10° rotation followed by resizing to 512×512 .
- Attack 7.* 250 pixel quarter cropping and a 60 pixel surrounded cropping.
- Attack 8.* 2 unit shearing transform in x and y directions.
- Attack 9.* The blind pattern matching attack (BPM) [4].

Tables 1 and 2 show the NC of the proposed method for Lena under the above-mentioned nongeometric and geometric attacks. As shown in these two Tables the average NC between the retrieved watermark and the original logo is 0.96. Also, according to the retrieved watermark logos in Table 1 and 2, the proposed scheme can retrieve clear and recognizable digital logos from the attacked images.

Moreover, the average NC of [4], [5], [9] and the proposed scheme for different test images are computed. Unlike other related schemes that did not measure the performance of their algorithms on any databases, we uses a database to evaluate the

Table 1. The performance of the proposed method by calculating the NC value and the retrieved logo under mentioned non geometric attacks in section 4.2











	Blurring	JPEG	Sharpening	Noise addition	BPM
Attacked Image					
PSNR(dB)	33.21	31.89	23.47	18.31	29.58
Retrieved Logo					
NC	0.993	0.995	0.99	0.973	0.998

Table 2. The performance of the proposed method by calculating the NC value and the retrieved logo under mentioned geometric attacks in section 4.2



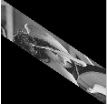







	Cropping (surround)	Cropping (quarter)	Shear X	Scaling	Rotation
Attacked Image					
PSNR(dB)	8.3	10.64	8.1	29.49	10.5
Retrieved Logo					
NC	0.846	0.824	0.99	0.992	0.998

Table 3. Comparing NC between the proposed schemes in [4], [5], [9] and our method. The result is averaged over twelve randomly selected images with different complexities from Washington University database under the mentioned attacks in section 4.2.

	Attack	Scheme [4]	Scheme [5]	Scheme [9]	Proposed scheme
Geometric Attacks	Rotation	0.65	0.68	0.92	0.99
	Shearing in x direction	0.49	0.49	0.58	0.98
	Shearing in y direction	0.51	0.51	0.63	0.99
	Surround cropping	0.53	0.51	0.72	0.83
	Quarter cropping	0.73	0.72	0.99	0.78
	Scaling	0.98	0.99	1	0.99
Non Geometric Attacks	Blurring	0.99	0.99	1	0.99
	Noising	0.98	0.99	1	0.97
	JPEG	0.99	0.99	1	0.99
	Sharpening	0.98	0.98	1	0.98
	BPM	0.97	0.98	1	0.99
	Average	0.80	0.80	0.89	0.95

robustness of our proposed method. As the robustness of each scheme is dependent on the complexity of test image, we classified images of Washington University database [15] into low, medium and high complexity based on quad tree representation [14]. Then we randomly selected four images from each complexity as test images and we simulated our experiments on these 12 test images.

We applied the above-mentioned attacks on each 12 test images. Table 3 summarizes the average NC under above attacks over 12 test images in schemes [4], [5], [9] and our method. As indicated in Table 3, for geometric attacks our scheme outperforms other schemes in most cases such as rotation and shearing. For non geometric attacks although other schemes have better performance than ours, the differences in NC values are small. However, Table 3 shows that our scheme gives higher average NC in comparison with other schemes.

5 Conclusion

As zero-watermarking algorithms do not degrade the quality of the original images, it has become a new research area in recent years. These methods extract binary features from the host image and then apply exclusive-or between the binary features and a binary logo to generate a verification map. The resulted verification map is registered in a trusted authority for further protection. One of the most important issues in copyright protection schemes is how to retrieve the significant binary features from an image that is aimed to be watermarked. In this paper we used discrete wavelet transformation, cellular automaton and an effective normalization procedure to extract binary features from an image. The normalization procedure is used to provide robustness against geometric distortions of the image. Also, a cellular automaton is utilized to maintain the robustness against noises.

Experimental results demonstrate that the proposed scheme has high performance against common geometric attacks and sufficient robustness under non geometric attacks. Consequently, according to the simulation results on images with different complexities our proposed method outperforms other similar schemes [4], [5] and [9] in most cases.

References

1. Langelaar, G., Setyawan, I., Lagendijk, R.: Watermarking digital image and video data: A state-of-the-art overview. *IEEE Signal Processing Magazine* 17(5), 20–46 (2000)
2. Lee, W.B., Chen, T.H.: A public verifiable copy protection technique for still images. *J. Syst. Softw.* 62, 195–204 (2002)
3. Hsu, C.T., Wu, J.L.: Multiresolution watermarking for digital images. *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.* 45(8), 1097–1101 (1998)
4. Chen, T.H., Horng, G., Lee, W.B.: A publicly verifiable copyright proving scheme resistant to malicious attacks. *IEEE Transactions on Industrial Electronics* 52(1), 327–334 (2005)

5. Abdel-Wahab, M.A., Selim, H., Sayed, U.: A novel robust watermarking scheme for copyright-proving . In: The 2009 International Conference on Computer Engineering and Systems, ICCES 2009, art. no. 5383216, pp. 482–486 (2009)
6. Chang, C.C., Lin, P.Y.: Adaptive watermark mechanism for rightful ownership protection. *The Journal of Systems and Software* 81, 1118–1129 (2008)
7. Kazakova, N., Margala, M., Durdle, N.G.: Sobel edge detection processor for a real-time volume rendering system. In: *Proceedings of the 2004 International Symposium on Circuits and Systems*, vol. 2, pp. 913–916 (2004)
8. Lee, M.T., Chen, S.S.: Image Copyright Protection Scheme Using Sobel Technology and Genetic Algorithm. In: *International Symposium on Computer, Communication, Control and Automation* (2010)
9. Lin, W.H., Horng, S.J., Kao, T.W., Chen, R.J., Chen, Y.H.: Image copyright protection with forward error correction. *Expert Systems with Applications* 14(12) (2005)
10. Popovici, A., Popovici, D.: Cellular automata in image processing. In: *Proceedings of 15th International Symposium on Mathematical Theory of Networks and Systems* (2002)
11. Dong, P., Brankov, J.G., Galatsanos, N.P., Yang, Y., Davoine, F.: Digital Watermarking Robust to Geometric Distortions. *IEEE Transaction on Image Processing* 36, 11888–11894 (2009)
12. Rothe, I., Susse, H., Voss, K.: The method of normalization to determine invariants. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 18(4), 366–376 (1996)
13. Hsu, C.T., Wu, J.L.: Hidden digital watermarks in images. *IEEE Transaction on Image Processing* 8(1), 58–68 (1999)
14. Yaghmaee, F., Jamzad, M.: Computing watermark capacity in images according to their quad tree. In: *Int. IEEE 5th Symposium on Signal Processing and Information Technology*, pp. 823–826. IEEE Press (2005)
15. Image data base,
<http://www.cs.washington.edu/research/imagedatabase>

Multi-view Video Coding Based on High Efficiency Video Coding

Kwan-Jung Oh, Jaejoon Lee, and Du-Sik Park

Advanced Media Lab, Samsung Electronics Co., Ltd
San 14, Nongseo-dong, Giheung-gu, Yongin-si, Gyeonggi-do, 446-712 Korea
{kwanjung.oh, jaejoon1.lee, dusikpark}@samsung.com

Abstract. Multiview video coding is one of the key techniques to realize the 3D video system. MPEG started a standardization activity on 3DVC (3D video coding) in 2007. 3DVC is based on multiview video coding. MPEG finalized the standard for multiview video coding (MVC) based on H.264/AVC in 2008. However, High Efficiency Video Coding (HEVC) which is a 2D video coding standard under developing outperforms the MVC although it does not employ interview prediction. Thus, we designed a new multiview video coding method based on HEVC. Interview prediction was added into HEVC and some coding tools were refined to be proper to MVC. The encoded multiple bitstreams are assembled into one bitstream and it is decoded into multiview video at decoder. From experimental results, we confirmed that the proposed MVC based on HEVC is much better than H.264/AVC, MVC, and HEVC. It achieves about 59.95% bit saving compared to JMVC simulcast at the same quality.

Keywords: Multiview video coding, High efficiency video coding, HEVC, 3D video coding, 3DVC.

1 Introduction

The successive development of multimedia systems and network has been contributed to the standardization of video compression. In the last three decades, several international video coding standards have been established, for example, H.261/H.263 for video telephony and MPEG-1 and MPEG-2 for video CD and digital TV [1], [2]. After that, the MPEG-4 part 2, object oriented video coding standard, and the most popular video coding standard, H.264/AVC (Advanced Video Coding), were produced [3]. Although several further extensions such as FExt (Fidelity Range Extension) [4] and SVC (Scalable Video Coding) [5] were added to the H.264/AVC, many researchers have desired to develop a new video coding standard more efficient than H.264/AVC. A formal joint Call for Proposals (CfP) [6] on a next-generation video coding technology was issued in January 2010 by ITU-T VCEG (video coding experts group) and ISO/IEC MPEG (moving picture experts group), and proposals were evaluated at the first meeting of the MPEG & VCEG, JCV-VC (Joint Collaborative Team on Video Coding) [7], which took place in April 2010.

With the success of three-dimensional (3D) movies and various 3D display devices, several international 3D video coding scenarios also have been standardized with the 2D video coding standards. MPEG-2 MVP (multiview profile) [8], MPEG-4 MAC (multiple auxiliary component) [9], and H.264/AVC MVC (multiview video coding) was standardized for multiview video coding. MVC was standardized in July 2008 [10] and currently MVC suitability for interlaced multiview video is investigated. MPEG started a standardization activity on 3DVC (3D video coding) in 2007. The MPEG 3DVC announced a version on 3D video which supports various 3D displays and its new coding standard will be developed in two years [11]. Related works on 3D video coding include depth coding, virtual view synthesis (rendering), depth estimation as well as multiview video coding. In April 2011, a CfP on 3DVC was issued and various 3D video coding technologies have been studied.

In this paper, we propose a multiview video coding based on high efficiency video coding. We believe that the proposed approach will be the basis of the 3D video coding standard since MVC and HEVC are the most powerful coding standards for multiview video coding and 2D video coding. The rest of this paper is organized as follows. In Section 2, HEVC is introduced. Implementation of the proposed method is explained in Section 3. In Section 4, experimental results are given. Section 5 presents conclusions.

2 High Efficiency Video Coding

HEVC [12] is a new video coding standard and this is currently under developing in JCT-VC established by ISO/IEC MPEG and ITU-T VCEG. It aims to substantially improve coding efficiency compared to H.264/AVC and is targeted at next-generation HDTV (Ultra HDTV, 7680x4320) displays as well as improved visual quality in terms of noise, color gamut, and high dynamic range [13]. The first formal HEVC test model, HM, was established in the third JCT-VC meeting held in Guangzhou, China and it provides more flexibility than H.264/AVC. The basic coding unit (CU) has a similar role to the macroblock in H.264/AVC. CU can be further split into prediction unit (PU). Transform unit (TU) is defined for transform and quantization. The overall coding structure is characterized by various sizes of CU, PU, and TU in a recursive manner, once the size of the largest coding unit (LCU) and the hierarchical depth of CU are defined [13].

The CU allows content-adaptive recursive splitting into four equally sized blocks, starting from 64x64 to 8x8 for luma samples. Both skipped CU and non-skipped CU types are allowed. The skipped CU is considered to be an inter prediction mode without coding of motion vector differences and residual data. The non-skipped CU is assigned to one of two prediction modes, intra and inter. Fig. 1 shows the example of CU structure.

The PU is the elementary unit used for carrying the information related to the prediction. In general, it is not restricted to being square in shape, in order to facilitate partitioning which matches the boundaries of real objects in the picture. Each CU may contain one or more PUs. PU types can be skip, intra, and inter. Fig. 2 shows the four types of PU structure.

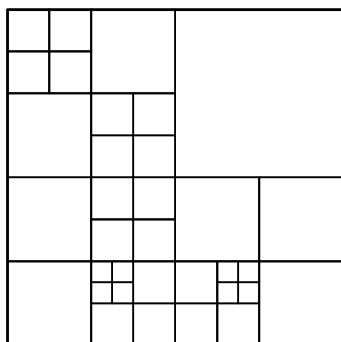


Fig. 1. Example of Coding Unit Structure

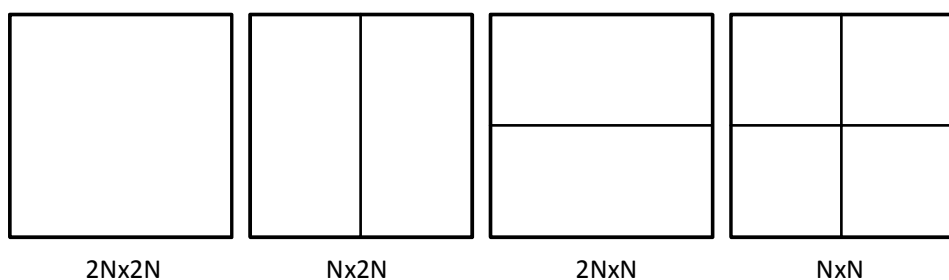


Fig. 2. Four Types of Prediction Unit Structure

The third type of unit is the TU, which is the unit for transform and quantization. It should be noted that the size of TU may be larger than that of the PU but not exceed that of the CU. It is always square and it may take a size from 4×4 up to 32×32 luma samples. Each CU may contain one or more TUs, where multiple TUs may be arranged in a quadtree structure, as illustrated in Fig. 3. Fig. 4 shows the relationship between CU, PU, and TU.

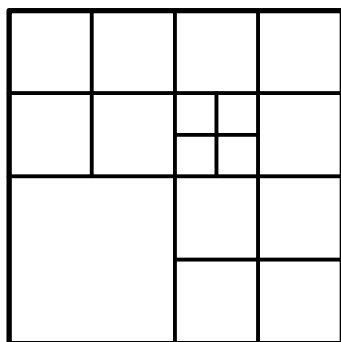


Fig. 3. Example of Transform Unit Structure

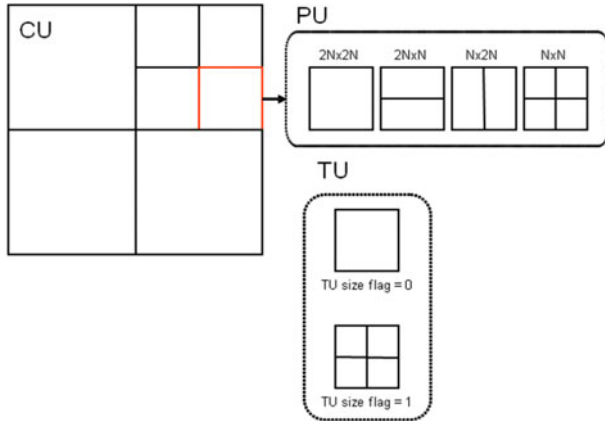


Fig. 4. Relationship between CU, PU, and TU

In addition to new coding structure, HEVC supports various new techniques such as, angular intra prediction, AMVP (advanced motion vector prediction), IBDI (internal bit depth increase), variable size transform/quantization, ALF (adaptive loop filter), and etc [14]. Currently, two configurations are suggested for typical coding tools. They are for high efficiency and low complexity, respectively as in Table. 1.

Table 1. Structure of tools in HM configurations

High Efficiency Configuration	Low Complexity Configuration
Coding Unit tree structure (8x8 up to 64x64 luma samples)	
Prediction Units	
Transform unit tree structure (3 level max.)	Transform unit tree structure (2 level max.)
Transform block size of 4x4 to 32x32 samples (always square)	
Angular Intra Prediction (34 directions max.)	
DCT-based interpolation filter for luma samples (1/4-sample, 8-tap)	
DCT-based interpolation filter for luma samples (1/8-sample, 4-tap)	
Coding Unit based Skip & Prediction Unit based merging	
Advanced motion vector prediction	
CABAC	Low complexity entropy coding phase 2
Internal bit-depth increase (2 bits)	X
X	Transform precision extension (2 bits)
Deblocking filter	
Adaptive loop filter	X

3 Implementation Multiview Video Coding Using HEVC

H.264/AVC MVC manages two reference software JMVM (joint multiview video model) [15] and JMVC (joint multiview video coding) [16] where several tools in JMVM are removed, such as illumination compensation, motion skip mode, and some tools related to SVC. However, the basic coding architectures of JMVM and JMVC are the same. We implement the interview prediction in JMVC to HEVC. For that, MVC related syntaxes, sequence parameter set MVC extension syntax and NAL unit header MVC extension syntax, are added to HEVC. Above two syntaxes are defined in Table 2 and Table 3. The proposed MVC based on HEVC supports hierarchical-B coding structure in temporal direction and I-B-P suture for view direction. The basic view-temporal coding structure is same with that of the JMVC as shown in Fig. 5.

Table 2. Sequence parameter set MVC extension syntax

Syntax	C	Descriptor
num_views_minus1	0	ue(v)
for(i = 0; i <= num_views_minus1; i++)		
view_id[i]	0	ue(v)
for(i = 1; i <= num_views_minus1; i++) {		
num_anchor_refs_10[i]	0	ue(v)
for(j = 0; j < num_anchor_refs_10[i]; j++)		
anchor_ref_10[i][j]	0	ue(v)
num_anchor_refs_11[i]	0	ue(v)
for(j = 0; j < num_anchor_refs_11[i]; j++)		
anchor_ref_11[i][j]	0	ue(v)
}		
for(i = 1; i <= num_views_minus1; i++) {		
num_non_anchor_refs_10[i]	0	ue(v)
for(j = 0; j < num_non_anchor_refs_10[i]; j++)		
non_anchor_ref_10[i][j]	0	ue(v)
num_non_anchor_refs_11[i]	0	ue(v)
for(j = 0; j < num_non_anchor_refs_11[i]; j++)		
non_anchor_ref_11[i][j]	0	ue(v)
}		

Table 3. NAL unit header MVC extension syntax

Syntax	C	Descriptor
nal_unit_header_mvc_extension() {		
non_idr_flag	All	u(1)
priority_id	All	u(6)
view_id	All	u(10)
temporal_id	All	u(3)
anchor_pic_flag	All	u(1)
inter_view_flag	All	u(1)
reserved_one_bit	All	u(1)
}		

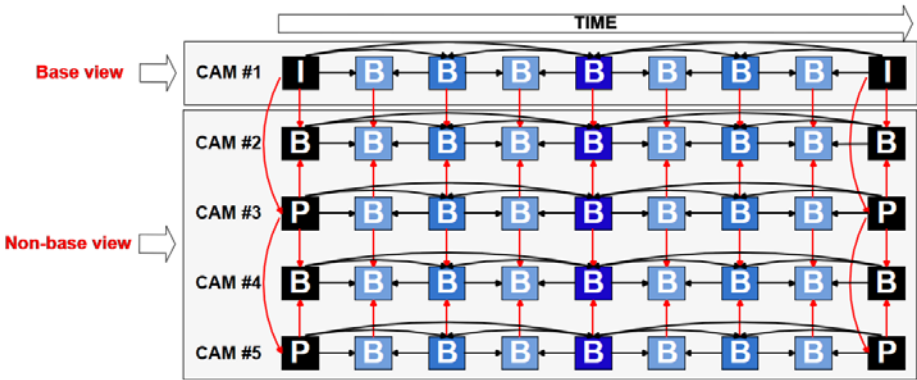


Fig. 5. View-temporal Coding Structure for MVC

For the base view in Fig. 5, we encode the sequence parameter set MVC extension syntax as prefix for slice header as show in Fig. 6. The base view is compatible with HEVC by skipping white colored NAL units such as Subset SPS, additional PPS, and prefixes.

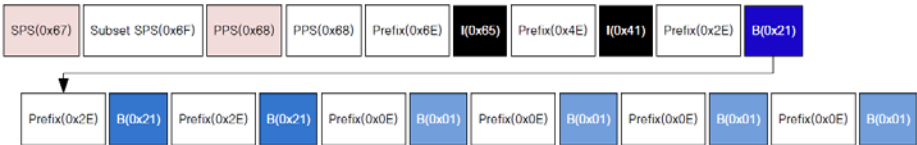


Fig. 6. Arrangement of NAL Units for Base View

For the non-base views, a new slice header (0x14) containing a prefix (0x2E) information and slice header information (0x01) is defined. Thus, it does not need to send a prefix defined in base view. Fig. 7 shows an arrangement of NAL units for P view and B view.

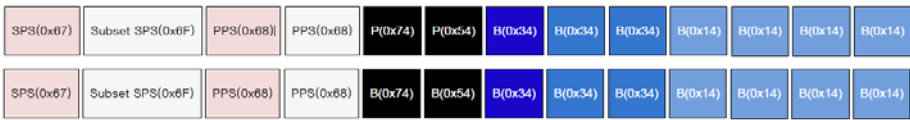


Fig. 7. Arrangement of NAL Units for Non-base View

Multiple encoded bitstreams are unified one bitstream by assembler. Fig. 8 shows the NAL units arrangement of assembled bitstream in case of five views. Fig. 9, Fig. 10, and Fig. 11 show how encoder, assembler, and decoder are working.

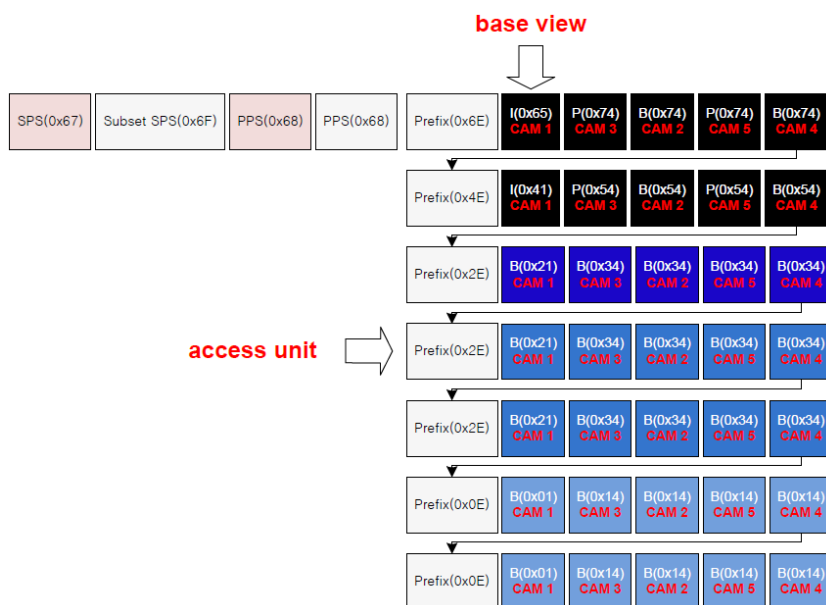


Fig. 8. NAL Unit Arrangement in Assembled Bitstream

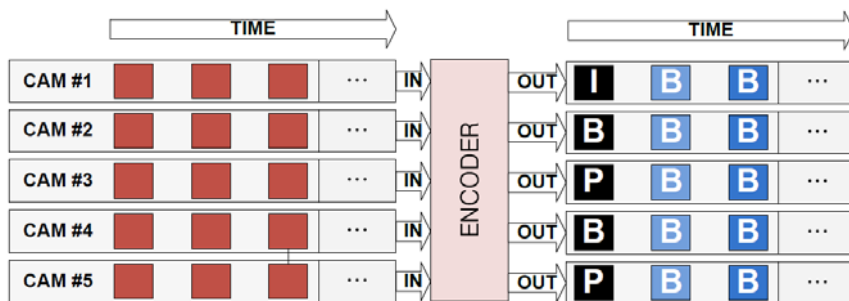


Fig. 9. Encoder

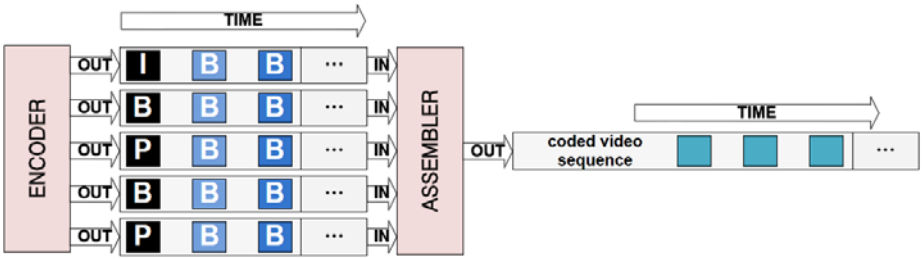


Fig. 10. Bitstream Assembler

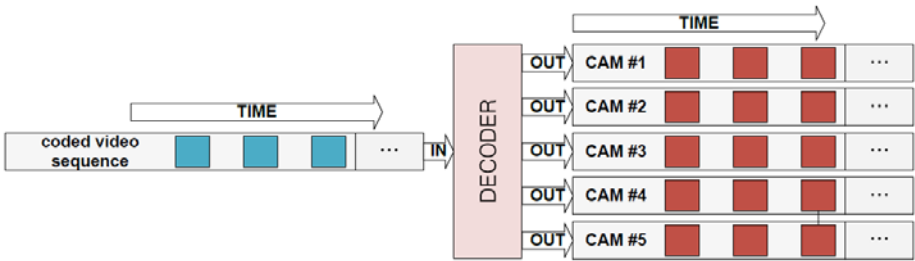


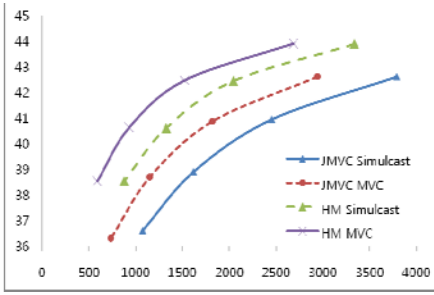
Fig. 11. Decoder

4 Experimental Results

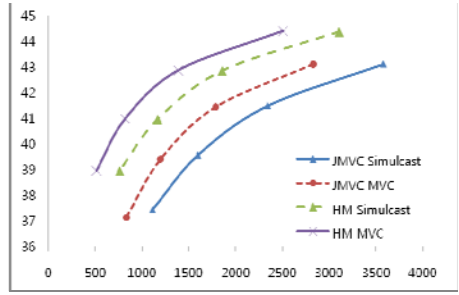
The proposed multiview video coding using high efficiency video coding was implemented based on HM 2.0. To verify the efficiency of the proposed method, we tested JMVC, HM simulcast, and proposed method compared to JMVC simulcast. Encoding of 61 frames for each view with QP 24, 28, 32, 36 was performed for 3 views case with the same test conditions as in MPEG 3DVC CfP [17]. The efficiency of the proposed algorithm was evaluated with BDBR (Bjontegaard delta bit rate) metric [18]. Table 4 shows the experimental results and Fig. 12 shows RD curves.

Table 4. Experimental results in terms of BDBR

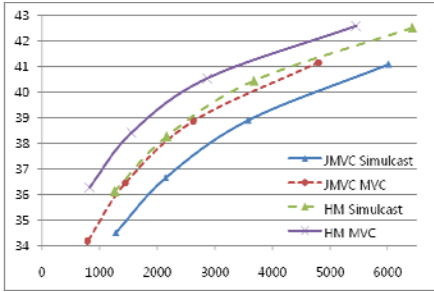
Sequence	JMVC	HM Simulcast	Proposed (HM MVC)
Balloons	-25.38 %	-41.99 %	-59.37 %
Kendo	-22.43 %	-44.47 %	-60.92 %
Lovebird1	-27.40 %	-30.33 %	-50.96 %
Newspaper	-18.97 %	-36.27 %	-52.38 %
GT_Fly	-39.03 %	-30.02 %	-68.57 %
Poznanhall2	-12.89 %	-54.96 %	-66.54 %
Poznanstreet	-35.56 %	-34.62 %	-58.34 %
Undo_Dancer	-40.12 %	-30.02 %	-61.95 %
Average	-28.06 %	-37.24 %	-59.95 %



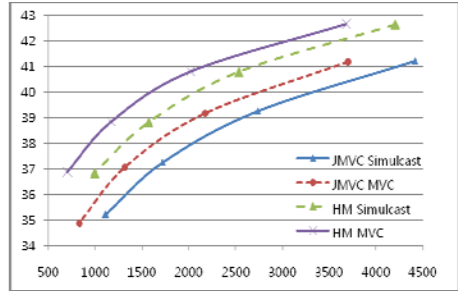
(a) Balloons



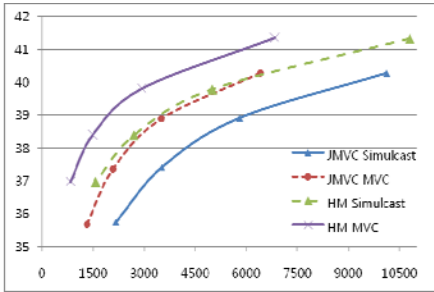
(b) Kendo



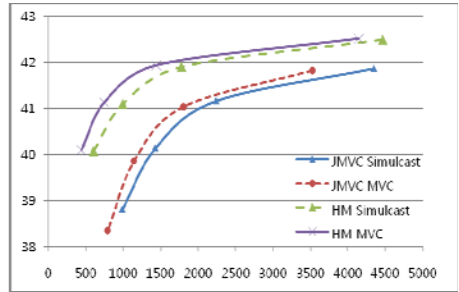
(c) Lovebird1



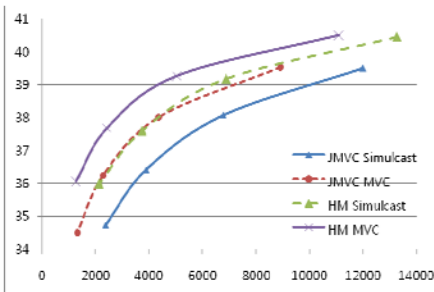
(d) Newspaper



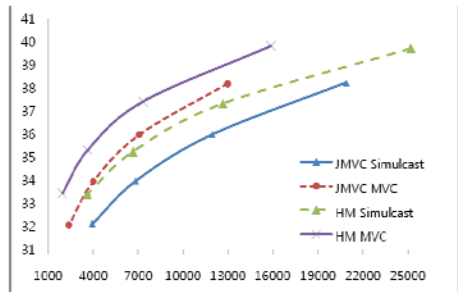
(e) GT_Fly



(f) Poznanhall2



(g) Poznanstreet



(h) Undo_Dancer

Fig. 12. RD Curves

5 Conclusions

In this paper, multiview video coding using high efficiency video coding is proposed. Interview prediction based on JMVC was added to HM. In addition, an assembler to combine multiview bitstreams into one bitstream was implemented. To verify the efficiency of the proposed method, four coding approaches were tested; JMVC simulcast, JMVC, HM simulcast, and the proposed method. From the experimental results, we confirmed that the proposed HM MVC outperforms the other approaches. It achieves about 59.95% bit saving compared to JMVC simulcast at the same quality. We expect that the proposed method will be a basement of MPEG 3DVC standard.

References

1. Hang, H.M., Woods, J.W.: Handbook for Visual Communications. Academic Press (1995)
2. Rao, K.R., Hwang, J.J.: Techniques and for Image, Video, and Audio Coding. Prentice-Hall (1996)
3. Richardson, I.E.: H.264 and MPEG-4 Video Compression: Video Coding for Next Generation Multimedia. Wiley (2003)
4. Sullivan, G.J., Topiwala, P., Luthra, A.: The H.264/AVC Advanced Video Coding Standard: Overview and Introduction to the Fidelity Range Extensions. In: SPIE Conference on Applications of Digital Image Processing, vol. 5558(1), pp. 454–474 (2004)
5. Schwarz, H., Marpe, D., Wiegand, T.: Overview of the Scalable Video Coding Extension of the H.264/AVC Standard. IEEE Trans. CSVT 17(9), 110–120 (2007)
6. ITU-T VCEG and ISO/IEC MPEG VCEG-AM91 & N11113: Joint Call for Proposals on Video Compression Technology (2010)
7. ITU-T VCEG and ISO/IEC MPEG VCEG-AM90 & N11112: Terms of Reference of the Joint Collaborative Team on Video Coding Standard Development (2010)
8. Chen, X., Luthra, A.: MPEG-2 multiview profile and its application in 3D TV. In: SPIE-Multimedia Hardware Architectures, vol. 3021, pp. 212–223 (1997)
9. Karim, H.A., Worrall, S., Sadka, A.H., Konoz, A.M.: 3-D video compression using MPEG-4-multiple auxiliary component (MPEG4- MAC). In: IEE 2nd International Conference on Visual Information Engineering (2005)
10. ISO/IEC JTC1/SC29/WG11 w9978: Text of ISO/IEC 14496-10:2008/FDAM 1 Multiview Video Coding (2008)
11. ISO/IEC JTC1/SC29/WG11 N10357: Vision on 3D Video (2009)
12. Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11 JCTVC-A124: Video coding technology proposal by Samsung (and BBC) (2010)
13. High efficiency video coding, <http://www.h265.net>
14. ITU TSB: Joint Collaborative Team on Video Coding. ITU-T Retrieved 21 (2010)
15. ISO/IEC JTC1/SC29/WG11 JVT-AA207: Joint multi-view video model (JMVM) 8.0 (2008)
16. ISO/IEC JTC1/SC29/WG11 JVT-AD207: WD 4 reference software for MVC (2009)
17. ISO/IEC JTC1/SC29/WG11 N12036: Call for Proposals on 3D Video Coding Technology (2011)
18. ITU-T SG16 Q.6 VCEG-AE07: An Excel Add-in for Computing Bjøntegaard Metric and its Evolution (2007)

2D to 3D Image Conversion Based on Classification of Background Depth Profiles

Guo-Shiang Lin¹, Han-Wen Liu², Wei-Chih Chen²,
Wen-Nung Lie², and Sheng-Yen Huang³

¹Dept. of Computer Science and Information Engineering, Da-Yeh University
168, University Rd., Dacun, Chang-Hua, 515, Taiwan, R.O.C.

khlin@mail.dyu.edu.tw

²Department of Electrical Engineering
National Chung Cheng University, Chia-Yi, 621, Taiwan, R.O.C.

ieewn1@ccu.edu.tw

³Reallusion Inc., New Taipei City, Taiwan
elvis@reallusion.com.tw

Abstract. In this paper, a 2D to 3D stereo image conversion scheme is proposed for 3D content creation. The difficulty in this problem lies on depth estimation/assignment from a mono image, which actually does not have sufficient information. To estimate the depth map, we adopt a strategy of performing foreground/background separation first, then classifying a background depth profile by neural network, estimating foreground depth from image cues, and finally combining them. To enhance stereoscopic perception for the synthesized images viewed on 3D display, depth refinement based on bilateral filter and HVS-based contrast modification between the foreground and background are adopted. Subjective experiments show that the stereo images generated by using the proposed scheme can provide good 3D perception.

Keywords: 2D to 3D image conversion, background depth profile, stereoscopic perception, depth cue estimation.

1 Introduction

3D video applications, such as 3D multimedia, 3DTV broadcasting, and 3D gaming, are getting more popular due to an incredible viewing experience compared with 2D video. Among them, the 3D digital frame is promising in near future's consumer electronics products. Nowadays in the market, its LCD panel has been manufactured in a size of 7 inches that can be viewed without glasses (i.e., naked eye). A traditional 2D color image, either raw or decoded data, can then be converted into a left-and-right or a multi-channel format so as to be viewed on 3D displays.

To have a capability of multi-view conversion, the depth information that originally does not exist in the 2D color images needs to be estimated. Then, the Depth Image

Based Rendering (DIBR) technique can be used to render/synthesize stereo or multi-views. Currently, researchers have proposed several 2D to 3D conversion algorithms for static images [2,13-15,17] and dynamic videos [4-6,12], aiming to mitigate the insufficiency of 3D contents. Due to less depth cues (e.g., motion) that can be found compared to video, images' 2D to 3D conversion is much more challenging.

Recent researches about automatic depth estimation from 2D photographic images can be divided into two categories. The first one is depth from defocus/focus. S. K. Nayer and Y. Nakagawa [1] explored the relationship between the focus level and the object distance from the focused plane, called SFF (Shape from focus), to estimate the depths. This method demands multiple images captured with different focal lengths, which is beyond our discussion. V. P. Namboodiri and S. Chaudhuri [2] proposed a method to perceive the depth layers from a single defocused image, called DFD (Depth from defocus). They estimate the blurring degree of each pixel and use it for assigning the relative depth. The other category of 2D depth estimation is based on multiple depth cues. For example, in [13], Hough transform is used to detect the vanishing point as the geometric cue, by which an initial depth map can be constructed. The depth map is then refined based on the texture cues extracted from the image segmentation result. In [14], wavelets transform of luminance (Y) component is used to detect high frequency of the foreground objects. For pixels of high spatial frequency, the depth is assigned larger (i.e., nearer). Their method is however preferably applicable to close-up images. On the other hand, Liu et al. [15] excludes the computation of depth cues from texture, contrast, or motion vector, but adopts a semantic-based algorithm which analyzes each image into parts of sky, land, building, etc. and assigns depths according to the result of semantic classification. On the other hand, Philips company [3] analyzes the image content to fit a background depth model. Discrete Cosine Transforms (DCTs) of the horizontal and the vertical projection profiles are performed and then a classifier is used to determine a best fit model according to the transformed coefficients.

In view of the human visual system (HVS), 3D space extensity perceived by human beings is mainly contributed from a layered or structured background depth and the relative depth between the foreground and the background. Based on this concept, we propose in this paper a 2D to 3D image conversion algorithm that integrates the processes of foreground/background separation, relative depth estimation for foregrounds, classification of and combination with a structured background depth profile, and post processing. Note that our classification of background depth profile is based on features of local texture gradient and local edge direction, aiming to provide better classification than that based on DCT coefficients of projection profiles. Our scheme is more generic and then more suitable for the conversion of 2D images including indoor, outdoor, landscapes, portrayal, etc.

The remainder of this paper is organized as follows. Section II describes the proposed depth estimation algorithm. Section III elaborates details of post processing to enhance the perceived stereo quality. In Section IV, experiment results are given and finally Section V draws some conclusions.

2 Proposed Depth Estimation Algorithm

Our proposed image conversion algorithm is illustrated in Fig.1, which consists of two parts: depth estimation and post processing. First, a segmentation-based method is applied to extract the foregrounds. Then the foreground depth and the background depth profile are estimated separately; the former is based on multiple depth cue estimation, while the later is based on neural classification. To enhance the perceived depth on a stereoscopic display, the initial depth map is refined by using the color information (e.g., alignment of color and depth edges) and the relative contrast between the foreground and background regions are tuned based on HVS. Finally, the refined color and depth information are both used to synthesize the stereo image pair by depth image based rendering (DIBR) technique.

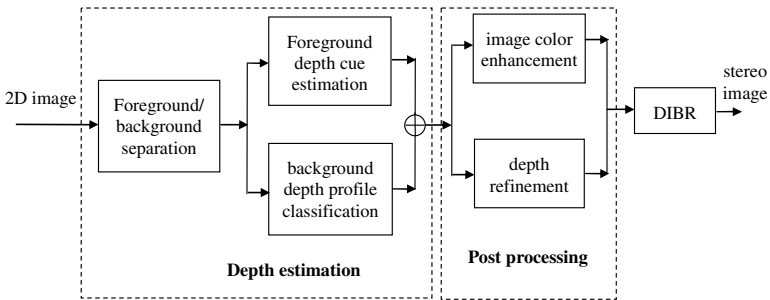


Fig. 1. The proposed 2D to 3D image conversion algorithm

2.1 Foreground/Background Separation

To extract the foreground regions, we adopt a strategy of performing region segmentation first and subsequently identifying regions that possibly belong to the foregrounds. There are several well-known region segmentation algorithms that have been proposed in literature. Among them, the mean-shift algorithm [10] is popularly used. Fig.2(b) demonstrates the segmentation result.

It is a challenging work to identifying foreground regions without some a priori knowledge. Based on an observation that foreground objects often occur at the central part of a frame (at least this assumption is valid for the digital frame application), we devise sampling boxes, as shown in Fig.2(b), to make statistics about the foreground (red) and background (green) color information. Since the pixel colors have been quantized by mean-shift algorithm, the results of color statistics will be limited. Denote the kinds of colors existing in the central and outer regions be $OC = \{oc_i | i = 1, \dots, M\}$ and $BC = \{bc_i | i = 1, \dots, N\}$, respectively. Our goal will be to delete from BC the colors that possibly belong to foregrounds. Colors that retain in the revised BC' will be used to extract the background regions.

Our method to classify the regions of a color bc_i in BC is to design filters based on a priori. A filter is used to sift out a color bc_i from BC if it also occurs in OC and satisfies a certain criterion according to some region features (note that bc_i may contain several disjointed regions in the frame). Region features are defined to include: position (x, y) and size $(\Delta x, \Delta y)$ of the smallest enclosing rectangle, and compactness (the ratio between the region area size and $\Delta x \cdot \Delta y$). The criteria used in filters may be, e.g., the bottom y of a background region should not be lower than a threshold (a lower region most probably belong to the foreground); regions of larger Δx are possibly foregrounds. Fig.2(c) shows an example of foreground extraction (i.e., classifying regions of colors in BC' as backgrounds and as foregrounds, otherwise). It can be seen that the result is satisfactory. Surely, the use of filters is not sufficient to sift out all false background colors in BC .

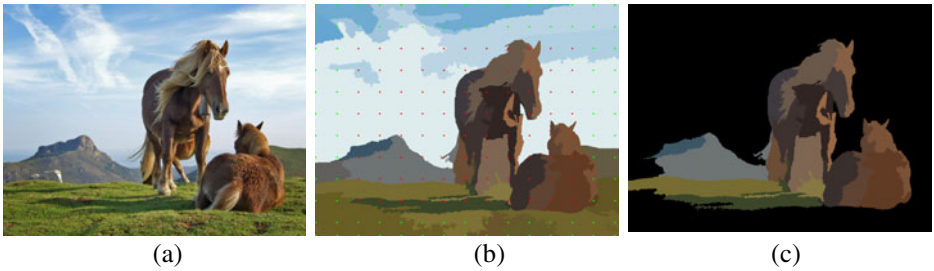


Fig. 2. (a) Original image, (b) result of mean-shift segmentation and the sampling boxes for the foregrounds (red) and the backgrounds (green), (c) identified foreground regions (shown with their mean-shift colors)

2.2 Foreground Depth Estimation

The cues for foreground depth estimation include texture gradient, sharpness, and face detection. Normally, a nearer object has stronger texture gradient and sharpness. However, these two cues are often indistinctive for the human face. We then adopt skin-color detection as an auxiliary tool to identify human faces as foregrounds. Note that since the depth image is usually smooth (i.e., of low spatial frequency), we calculate the depth cues based on blocks of 8×8 pixels to reduce the processing time.

(a) Texture gradient

The texture gradient of each pixel is calculated by using the Law’s eight masks [7]:

$$z_i(x, y) = \left| \sum_{k=1}^1 \sum_{l=1}^1 w_i(k, l) I(x+k, y+l) \right| \tag{1}$$

where $I(x, y)$ is the intensity value at position (x, y) , and $w_i(k, l)$, $i=1 \sim 8$, denote the Law’s masks. The texture gradient for each block is then defined as:

$$f^T(u, v) = \sum_{(x, y) \in \text{block}_{-(u, v)}} U \left(\sum_{i=1}^8 z_i(x, y) - T_{t1} \right) \tag{2}$$

where T_{t1} is a predefined threshold and (u, v) is the block index.

(b) Sharpness

Empirically, edges of a near object have a sharper contrast than those of a far object. We define the variance and contrast of the graylevel in each block as the cues:

$$f^V(u, v) = \frac{1}{63} \sum_{(x, y) \in \text{block}_{-(u, v)}} (I(x, y) - \bar{I}_{u, v})^2 \tag{3}$$

$$f^C(u, v) = \frac{I_{u, v}^{\max} - I_{u, v}^{\min}}{I_{u, v}^{\max} + I_{u, v}^{\min}} \tag{4}$$

where $\bar{I}_{u, v}$, $I_{u, v}^{\max}$, and $I_{u, v}^{\min}$ represent the average, maximum and minimum pixel values within the (u, v) -th block, respectively.

(c) Face cue

First, the input image is transformed from RGB to YCbCr color space. Pixels that satisfy conditions in both the RGB and YCbCr spaces are identified as the skin-color pixels [8]. Also, human’s hair [9] (black is assumed) can be detected by using the algorithm proposed in [9]. The skin-color and hair-color pixels are united to form the human’s information and assigned with a depth cue $f^p = 255$; otherwise $f^p = 0$.

(d) Depth cue fusion

Finally, the depth cues are fused to generate the depths for pixels located by the foreground mask obtained in Section 2.1 through Eq.(5):

$$f(u, v) = (w_1 \cdot f^V(u, v) + w_2 \cdot f^C(u, v) + w_3 \cdot f^T(u, v)) \cup f^p(u, v), \tag{5}$$

where $w_1 \sim w_3$ are predetermined weights ($w_1 + w_2 + w_3 = 1.0$) and “ \cup ” means pixel-wise maximum extraction. Since the depth cues are calculated in terms of blocks of 8×8 pixels, we apply a simple bilinear interpolation to rescale the foreground depth $f(u, v)$ to match the size of the input image.

2.3 Background Depth Profile Classification

We use a three-layer BPNN (Back-propagation Neural Network) to classify an image to one of the 5 types of background depth profiles. Figure 3 shows the 5 depth profiles defined in our system. The “1: up-bottom progressive” type is mostly often used in 2D-to-3D conversion. The “2: left-right progressive” and “3: right-left progressive”

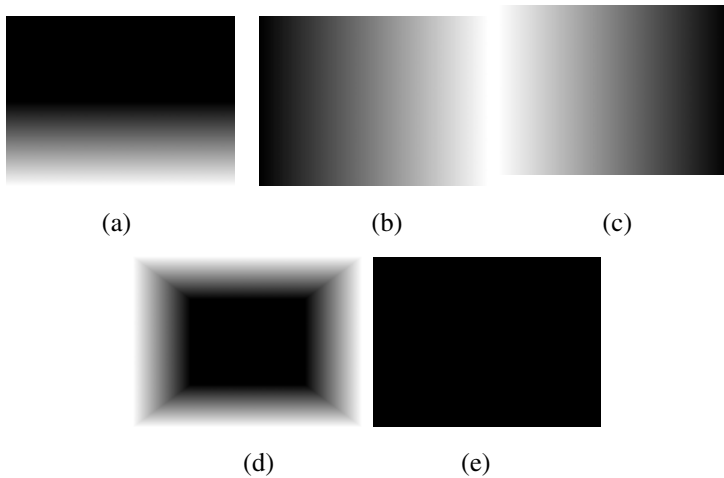


Fig. 3. Background depth profiles (a) up-bottom progressive, (b) left-right progressive, (c) right-left progressive, (d) indoor, and (e) close-up

show different increasing styles of depth. The “4: indoor” is most suitable for indoor scenario that constructs the strongest space extensity. On the other hand, “5: close-up” assumes the background depths all to zero, thus protruding the foreground objects substantially.

Features used in our neural network include:

1. local edge direction: The input image is divided into 3×3 regions, each is calculated the edge direction by using horizontal and vertical Sobel operator. All edge directions are quantized into 8 principle ones, each spaced by 45 degrees. The direction histograms of these 9 regions thus form the features of 72 dimensions.
2. local texture gradient: the average depth cues f^T in these 9 individual regions, calculated based on Eq.(2), are also adopted as features of 9 dimensions.

Our algorithm is based on a similar observation in [16] that edge directions will have a dominant pattern which can be used to calculate the vanish point. Hence features based on local edge direction and texture gradient will be much more promising in practice. The three-layer neural network carried out to classify the background depth profile of each input image have 81 ($72+9$) input neurons, 50 hidden-layer neurons, and 5 output neurons. All input features are first normalized to be between 0.0 and 1.0. The output neuron with the highest score is selected as the background depth profile.

2.4 Combination of Foreground and Background Depths

The foreground and background depths obtained above are individually normalized to (g_{\min}^F, g_{\max}^F) and (g_{\min}^B, g_{\max}^B) , respectively. These two sub-ranges can be fully or

partially overlapped (e.g., both are (0,255)), depending on user preferences. Finally, they are combined pixel-by-pixel by using a maximum operation. For non-foreground pixels, the final depth is the one calculated from the background depth profile; for foreground pixels, it is the maximum between foreground and background depths.

3 Color and Depth Post-processing

3.1 Depth Refinement

Since the depths are estimated at block-level first and then scaled up to the pixel-level, the depth edges may not be aligned with the color edges. This misalignment often causes quality degradation in the synthesized stereo images. We apply a bilateral filter [11] to refine the depth map. The bilateral filter is a weighted filter which evaluates the similarity of colors and distance between a current pixel and its neighboring ones, assigns the proper weights, and then calculates the weighted averages. It can not only smooth the depth map, but also make edges of foregrounds aligning to the color edges. Fig.4 (c) shows the refined depth map.

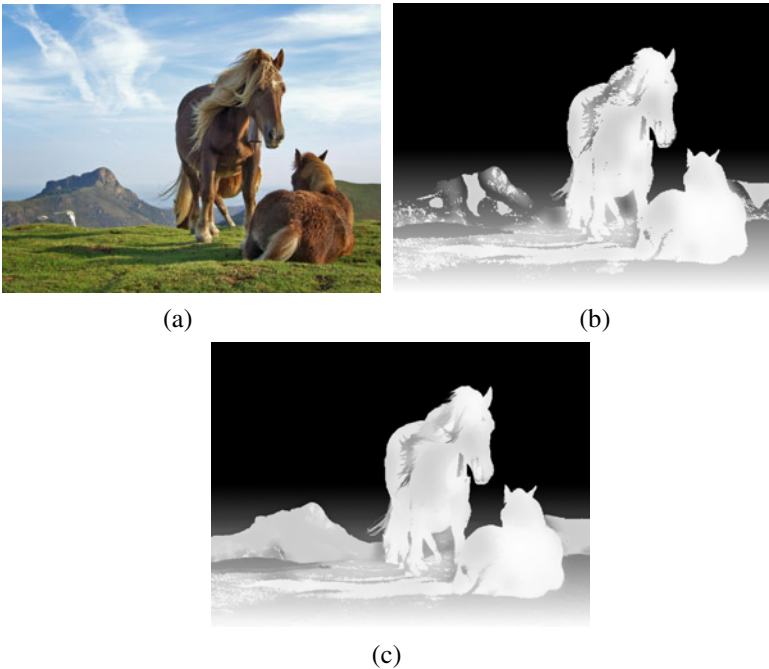


Fig. 4. Results of depth refinement (a) input image (b) initial depth estimation, (c) depth refined by bilateral filtering

3.2 Color Enhancement

It is known that the relative overlapping, foreground/background contrast, lighting, and shadows have influences on stereoscopic perception [12]. When looking at an image, people usually focus on foreground regions; the more the contrast between foreground and background regions, the more the stereoscopic perception. In this system, we apply two methods to modify colors of foreground/background pixels, according to the result of background depth classification, such that the stereoscopic effect is enhanced.

1. For background depth profiles #1-4, modify the RGB or hue-and-saturation (H/S) values of pixels in the *background* to increase its contrast w.r.t. the foreground;
2. for background depth profiles #5 (i.e., close-up), modify the RGB or hue-and-saturation (H/S) values of pixels in *foreground* to increase its contrast w.r.t. the background.

Figure 5 demonstrates the original images and the color-enhanced images, respectively. It is seen that with proper color enhancement, the space extensity can be enhanced.

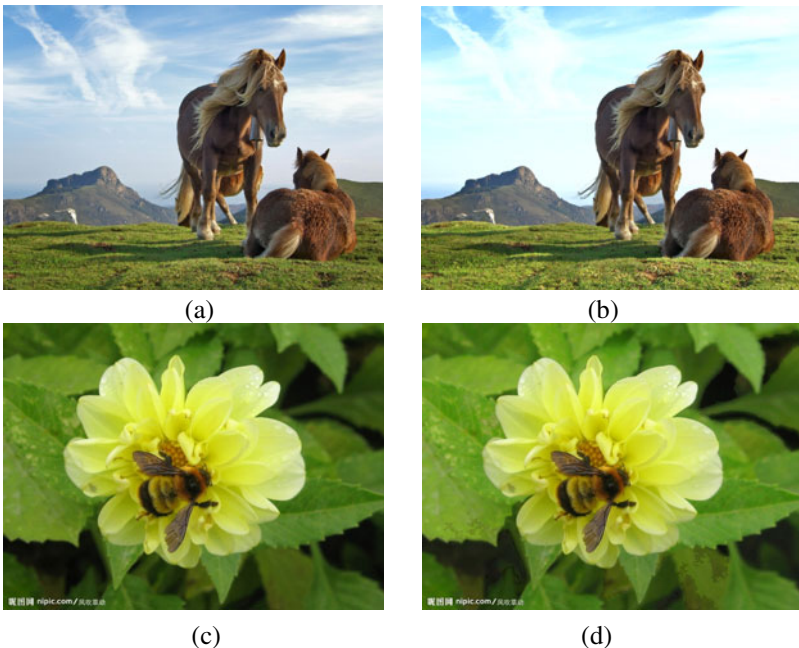


Fig. 5. (a)(c) Input images and (b)(d) color enhanced images

4 Experiment Results

The 3D display used in the experiments is an Acer 3D notebook (Model: 5738DG) equipped with polarizing glasses. Since depth estimation from a mono view is really challenging, we do not evaluate it by computing the objective quality metrics (e.g., PSNR) which require the existence of depth ground truths. Instead, subjective assessment based on Mean Opinion Score (MOS) of the synthesized stereo views is conducted. A total of 12 non-professional subjects are asked to score 1 to 5 (5 (excellent), 4 (good), 3 (fair), 2 (poor), and 1 (bad)) for each stereo pair generated by our proposed method (with $w_1=0.4$, $w_2=0.2$, and $w_3=0.4$). The test image size is all 640×480 pixels.

To evaluate background depth profile classification, 200 images (including landscape, portrait painting and indoor image, etc.) downloaded from the Internet are collected. Among them, the 5 types of background depth profiles are evenly distributed. A number of 75 images are used for training, 25 images for validation, and 100 images for testing. To determine the ground truths for neural network training, 5 subjects are asked to vote for the background depth profiles for each image. The dominant ones are selected as the truths. Table 1 shows the classification rates 91% and 83% for the training and test samples, respectively.

Table 1. Classification rates for background depth profiles

Type of depth profile	Classification rate (training sample)	Classification rate (test sample)
1	85 %	80 %
2	90 %	100 %
3	90 %	85 %
4	90 %	70 %
5	100 %	80 %
Average	91 %	83 %

Examples of some estimated depth maps are given in Fig. 6. Their background depth profiles are automatically classified as Type 1~5, respectively. The finally estimated depth maps are satisfactory.

Fig. 7 shows the MOS scores of the proposed algorithm for test images of different kinds of background depth profiles. Obviously, the close-up category yields the highest stereoscopic perception.

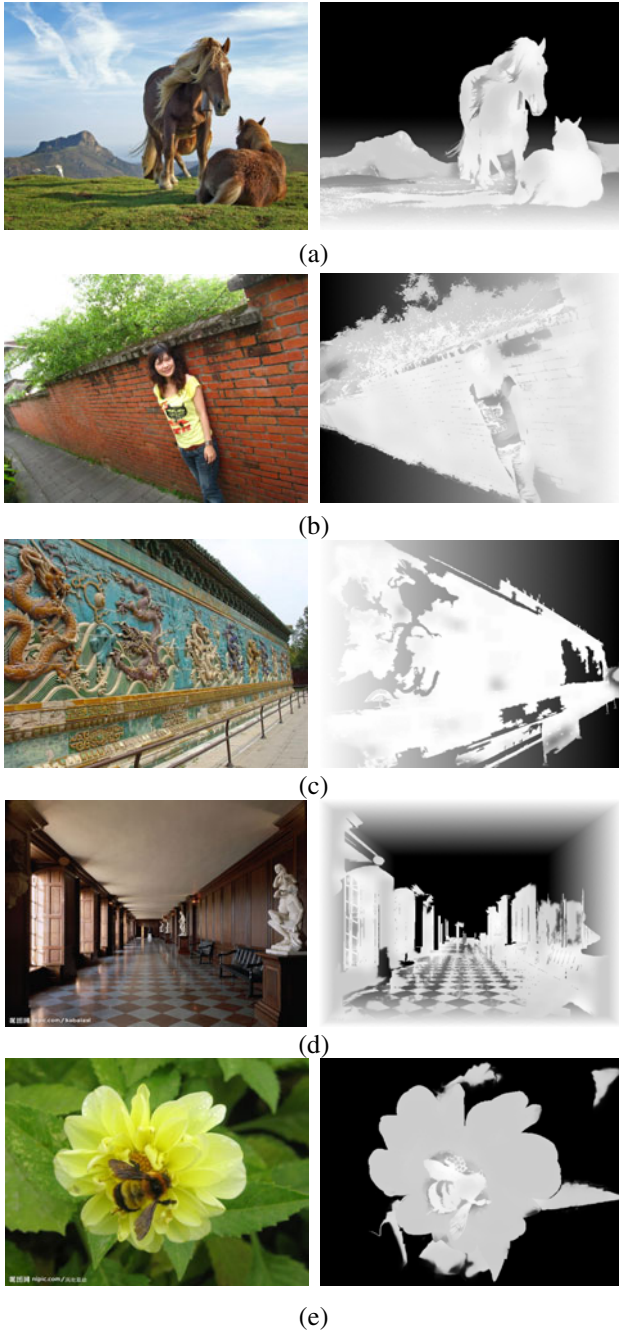


Fig. 6. (Left) input image (right) estimated depth map



Fig. 7. MOS scores for test images of different kinds of background depth profiles

5 Conclusions

In this paper, we propose a 2D to 3D image conversion scheme. Our scheme is featured of: 1) segmentation-based foreground extraction, 2) foreground depth estimation based on multiple depth cue, 3) neural-network-based background depth profile classification, and 4) color enhancement for stereoscopic perception. Experiments show that our background depth classification has achieved a correct rate of 83% and the quality of synthesized stereo images viewed on the 3D display is good.

References

1. Nayar, S.K., Nakagawa, Y.: Shape from Focus. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 16(8), 824–831 (1994)
2. Nambodiri, V.P., Chaudhuri, S.: Recovery of Relative Depth from a Single Observation Using an Uncalibrated (Real-Aperture) Camera. In: *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, Anchorage, AK, pp. 1–6 (2008)
3. Burazerovic, D., Vandewalle, P., Berretty, R.P.: Automatic Depth Profiling of 2D Cinema - and Photographic Images. In: *Proc. of IEEE International Conference on Image Processing*, Cairo, pp. 2365–2368 (2009)
4. Kim, M., Park, S., Kim, H., Artem, I.: Automatic conversion of two-dimensional video into stereoscopic video. In: *Proc. of SPIE*, vol. 6016, pp. 601610-1–601610-8 (2005)
5. Manbae, K., Sanghoon, P., Youngran, C.: Object-Based Stereoscopic Conversion of MPEG-4 Encoded Data. In: *Proc. of the 5th Pacific-Rim Conference on Multimedia*, pp. 491–498 (2004)
6. Kim, D., Min, D., Sohn, K.: A Stereoscopic Video Generation Method Using Stereoscopic Display Characterization and Motion Analysis. *IEEE Trans. on Broadcasting* 54(2), 188–197 (2008)

7. Suzuki, M.T., Yaginuma, Y., Yamada, T., Shimizu, Y.: A Shape Feature Extraction Method Based on 3D Convolution Masks. In: Proc. of Eighth IEEE International Symposium on Multimedia, pp. 837–844 (2006)
8. Peer, P., Kovace, J., Solina, F.: Human Skin Color Clustering for Face Detection. In: EUROCON 2003 International Conf. on Computer as a Tool, vol. 2, pp. 144–148 (2003)
9. Chen, Y.-J., Lin, Y.-C.: Simple Face-detection Algorithm Based on Minimum Facial Features. In: Proc. of IEEE International Conference on Industrial Electronics Society, pp. 455–460 (2007)
10. Sudhamani, M.V., Venugopal, C.R.: Segmentation of Color Images using Mean Shift Algorithm for Feature Extraction. In: Proc. of IEEE International Conference on Information Technology (2006)
11. Tomasi, C., Manduchi, R.: Bilateral Filtering for Gray and Color Images. In: Proc. of IEEE International Conference on Computer Vision, Bombay, pp. 839–846 (1998)
12. Lin, G.-S., Yeh, C.-Y., Chen, W.-C., Lie, W.-N.: A 2D to 3D conversion scheme based on depth cues analysis for MPEG videos. In: IEEE International Conference on Multimedia and Expo, pp. 1141–1145 (2010)
13. Han, K., Hong, K.: Geometric and texture cue based depth-map estimation for 2D to 3D image conversion. In: IEEE International Conference on Consumer Electronics, pp. 651–652 (2011)
14. Chiang, T.-W., Tsai, T., Lin, Y.-H., Hsiao, M.-J.: Fast 2D to 3D conversion based on wavelet analysis. In: IEEE International Conference on Systems Man and Cybernetics, pp. 3444–3448 (2010)
15. Liu, B., Gould, S., Koller, D.: Single image depth estimation from predicted semantic labels. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1253–1260 (2010)
16. Choi, J., Kim, W., Kong, H., Kim, C.: Real-time vanishing point detection using the local dominant orientation signature. In: 3DTV Conf., Turkey (May 2011)
17. Saxena, A., Sun, M., Ng, A.Y.: Learning 3-D Scene Structure from a Single Still Image. In: ICCV 2007 (2007)

Shape Matching and Recognition Using Group-Wised Points

Junwei Wang, Yu Zhou, Xiang Bai, and Wenyu Liu*

Department of Electronics and Information Engineering,
Huazhong University of Science and Technology, Wuhan, China
wjw.20012001@163.com, zhoyu.hust@gmail.com, xiang.bai@gmail.com,
liuwuy@hust.edu.cn

Abstract. Shape matching/recognition is a very critical problem in the field of computer vision, and a lot of descriptors and methods have been studied in the literature. However, based on predefined descriptors, most of current matching stages are accomplished by finding the optimal correspondence between every two contour points, i.e., in a pair-wised manner. In this paper, we provide a novel matching method which is to find the correspondence between groups of contour points. The points in the same group are adjacent to each other, resulting in a strong relationship among them. Two groups are considered to be matched when the two point sequences formed by the two groups lead to a perfect one-to-one mapping. The proposed **group-wised matching** method is able to obtain a more robust matching result, since the co-occurrence (order) information of the grouped points is used in the matching stage. We test our method on three famous benchmarks: MPEG-7 data set, Kimia's data set and Tari1000 data set. The retrieval results show that the new group-wised matching method is able to get encouraging improvements compared to some traditional pair-wised matching approaches.

Keywords: Shape matching, Pair-wised matching, Group-wised, Co-occurrence.

1 Introduction

Shape description and matching is a very critical problem in the field of computer vision. There are some important issues to be noted. First, geometric transformation invariance should be satisfied when matching two shapes, i.e., one shape and its translated/rotated/scaled versions are supposed to be very similar. Next, some intra-class variations, such as noise, articulation, local deformation and occlusion, should be carefully treated. The influences of these variances to shape similarity measure should be decreased to an acceptable level.

To handle the complex situations listed above, several shape representations and descriptors with "rich" shape information have been studied in the last decade, including Visual Part [2], Shape Context (SC) [3], Inner-distance Shape

* Corresponding author.

Context (IDSC) [9], Triangle Area Representation (TAR) [11], and Shape Tree [10]. Latecki and Lakamper introduced one novel shape representation called Visual Part [2]. Shapes are simplified by a process of digital curve evolution (DCE) and are further decomposed into perceptually meaningful parts, which are called *Visual Parts*. Shape matching is performed by looking for the optimal correspondence of visual parts. Shape Context [3], which utilizes the geometric relationship between contour sample points, is one of the most classical shape descriptors in the literature. For each sample point on one shape contour, its *Shape Context* captures the spatial distribution of all the other sample points relative to it, which offers a globally discriminative characterization. Ling and Jacobs proposed a novel distance definition called *inner distance* as a replacement for the usual Euclidean distance. The inner distance is defined as the length of the shortest path between landmark points within the shape silhouette, which is articulation insensitive and more effective at capturing part structures. Using the inner distance, Shape Context can be extended to a novel descriptor called Inner-Distance Shape Context [9]. Triangle Area Representation [11] presents a measure of convexity/concavity of each contour point using the signed areas of triangles formed by boundary points at different scales. The area value of triangle is a measure for the curvature of corresponding contour point. This representation is effective in capturing both local and global shape characteristics. Shape Tree [10] is one classical segment-based shape matching algorithm, which proposed a hierarchical representation for contour curves. The original curve can be broken into two halves by the middle point on it, and each of the two sub-curves can be broken into its halves. This hierarchical description can be represented by a binary tree, which is called the *Shape Tree* of a curve. The matching process is performed by comparing the hierarchical segments explicitly.

The methods introduced in the above paragraph are all effective shape descriptors. However, among all these methods, the matching stage is accomplished by finding the optimal correspondence between contour sample points or local parts. The sample points or shape parts, which are represented by the predefined descriptors, mainly consist of local characteristics of given shapes. For example, Shape Context, which is one of the most classical descriptors in recent years, pays more attention to local shape features. The bins in the histogram of Shape Context are uniform in a log-polar space, which makes it much more sensitive to nearby sample points than to points farther away [3]. Consequently, the contour sample points or parts are matched together probably because the two points or parts have similar local shape characteristics. This kind of matching strategy is sometimes not robust enough to deal with complex situations, when there are certain amount of inter/intra class variations.

In this paper, we present a novel method to achieve more robust shape matching. Our motivation comes from this observation: two contour points should take a correspondence, not only because these two points are very similar, but also the points related to them are very similar as well. If we take related points into consideration, the matching result is much more likely to be correct. Therefore, we put related points into one group, and when we perform matching between

two points, we consider matching their corresponding groups. We call this novel and robust matching strategy *group-wised matching*. In the next section, we will give some cues on how to judge which are related points and how to define groups.

A similar manner to our idea is the Dynamic Programming (DP) algorithm, which has been widely used for shape matching (e.g., by the method of Multi-scale Convexity Concavity (MCC) [6] and Inner-Distance Shape Context (IDSC) [9], etc.). In DP algorithm, the order information of contour sample points is utilized as a global constraint for optimal correspondence. However, in current DP algorithm, local point-wise misalignment is allowed, and the correspondence is still found between single points, not group of points. We combine our group-wised matching approach with the usual DP algorithm, and achieved better retrieval results (see experiments in Section 3).

The remainder of this paper is organized as follows. Section 2 introduces the proposed group-wised shape matching algorithm in details. Section 3 presents the experimental results. Finally, Section 4 makes some conclusions.

2 The Proposed Group-Wised Method

In this Section, we give the definition what is a group and the group-wised matching method in details.

2.1 Point Groups

Let $\mathbf{S} = \{\mathbf{p}_i\}$ ($i = 1, \dots, N$) denotes the sequence of equidistant sample points on the outer contour of a given shape \mathbf{S} , where the index i is according to the order of the sample points along the contour in the counter-clockwise direction. In our implementation, we set the number $N = 100$, which is consistent with the settings in many recent works, such as [6] and [9].

To build up the group-wised matching strategy, at first we need to give the definition of *group*. As stated in the previous section, it is a good choice to combine one given point and its related points on the shape contour to form a group. Then the problem is which point is related to the given point. We adopt one simple strategy here: the neighboring points are treated as the related points. That is because related points are more likely to get close to each other than get far away from each other. Just take the shape boundary of a human-being as an example. The contour points representing his/her “ears”, “eyes”, “nose” and “mouth” are closely related to each other, as all of them are the elements of the “head”. On the contrary, the contour points representing his/her “hands” and “feet” are almost unrelated to each other. Figure 1 gives another example of related points on the boundary of a horse shape. In Figure 1, some related points on the head of the horse are labeled as red dots connected by solid lines; some unrelated points on its legs and back are labeled as blue dots connected by dashed lines.

For the reason stated above, we choose points close to each other to form a group. Specifically, for one sample point \mathbf{p}_i , we choose k nearby points on

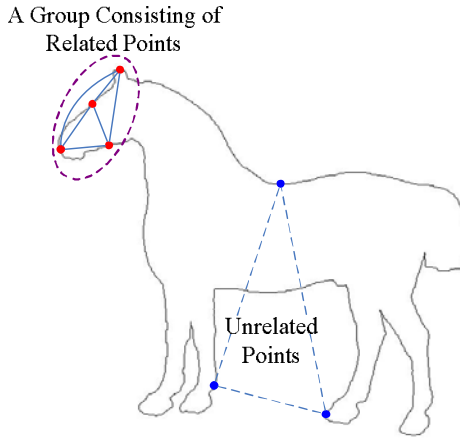


Fig. 1. An example of related/unrelated points. The “horse” shape is chosen from MPEG-7 CE-Shape-1 part B database [1]. See more details in the text.

the left hand side $\mathbf{p}_{i-1}, \mathbf{p}_{i-2}, \dots, \mathbf{p}_{i-k}$, and k nearby points on the right hand side $\mathbf{p}_{i+1}, \mathbf{p}_{i+2}, \dots, \mathbf{p}_{i+k}$. Note that the number of points on the left hand side is equivalent to the number of points on the right hand side, which leads to a balance description. With these $2k + 1$ points, we are able to define a meaningful point group, G_i . The point \mathbf{p}_i is right on the center of this group, which is regarded as the *centroid* of the group G_i . As there is co-occurrence (order) information among these contour sample points, the point group G_i is indeed a point subsequence belonging to the point sequence $\{\mathbf{p}_i\}$ ($i = 1, \dots, N$) on the outer contour of the shape \mathbf{S} :

$$G_i = (\mathbf{p}_{i-k}, \mathbf{p}_{i-k+1}, \dots, \mathbf{p}_{i-1}, \mathbf{p}_i, \mathbf{p}_{i+1}, \dots, \mathbf{p}_{i+k-1}, \mathbf{p}_{i+k}). \quad (1)$$

The parameter k , which controls the scale or the point number of the group, is regarded as the *radius* of the group. When k increases, the neighborhood of the point \mathbf{p}_i increases, and more points (with more shape characteristics) are included into the group.

For every sample point \mathbf{p}_i ($i = 1, \dots, N$), we define a point group G_i ($i = 1, \dots, N$). Then the whole shape \mathbf{S} is represented by a sequence of point groups:

$$\mathcal{G} = \mathcal{G}(\mathbf{S}) = (G_1, G_2, \dots, G_N). \quad (2)$$

2.2 Group-Wised Matching

For the task of shape recognition, usually a shape similarity or dissimilarity (distance) is computed by finding the optimal correspondence of contour points. Then the shape dissimilarity is the sum of the distances of the corresponding points. As defined in Section 2.1, every sample point is represented by one group consisting of itself and several neighboring points. Consequently, the shape

similarity is computed by finding the optimal correspondence of the predefined groups, and the shape dissimilarity value is the sum of the distances of the corresponding groups.

To match two shapes \mathbf{X}, \mathbf{Y} (represented by two sequences of point groups), the dissimilarity between any pair of points (indeed groups) should be computed. Let $\mathbf{p}_i, \mathbf{q}_j$ denote contour points of \mathbf{X}, \mathbf{Y} , respectively, and G_i, G_j denote corresponding groups for the two points. When we perform a matching between these two sample points $\mathbf{p}_i, \mathbf{q}_j$, we indeed perform a matching between the two groups G_i, G_j . To match the two groups, we consider not only the cost for mapping the two center points (i.e., the centroid of the groups) $\mathbf{p}_i, \mathbf{q}_j$, but also the costs for mapping their neighboring points within corresponding groups. Specifically, if the original pair-wised cost (feature distance) for mapping two points $\mathbf{p}_i, \mathbf{q}_j$ ($i, j = 1, \dots, N$) is denoted as $c(\mathbf{p}_i, \mathbf{q}_j)$, then our *group-wised matching* method defines the mapping cost for the two groups G_i, G_j as:

$$d(\mathbf{p}_i, \mathbf{q}_j) = d(G_i, G_j) = c(\mathbf{p}_i, \mathbf{q}_j) + \sum_{t=1}^k \omega_t \{c(\mathbf{p}_{i-t}, \mathbf{q}_{j-t}) + c(\mathbf{p}_{i+t}, \mathbf{q}_{j+t})\}. \quad (3)$$

In this way, shape matching is achieved through a group-wised manner. That is, not only the feature distance between the two points $\mathbf{p}_i, \mathbf{q}_j$ but also the feature distances between the neighboring points of the two points $\mathbf{p}_i, \mathbf{q}_j$ are utilized. Clearly, the pair-wised cost information $c(\mathbf{p}_i, \mathbf{q}_j)$ is completely included in the group-wised cost $d(\mathbf{p}_i, \mathbf{q}_j)$.

Note that the co-occurrence (order) information of every point group is perfectly included into the mapping cost defined in Equation 3, as we assume that the two groups G_i and G_j must have a perfect one-to-one mapping in their sequence point order, i.e., \mathbf{p}_{i-t} corresponds to \mathbf{q}_{j-t} and \mathbf{p}_{i+t} corresponds to \mathbf{q}_{j+t} for all $t = 1, \dots, k$. Figure 2 shows the idea of our group-wised matching method along with the difference between group-wised and pair-wised approaches.

This *group-wised mapping cost* can also be regarded as one novel pair-wised mapping cost for each pair of two points: the original pair-wised cost $c(\mathbf{p}_i, \mathbf{q}_j)$ is replaced by the novel pair-wised cost $d(\mathbf{p}_i, \mathbf{q}_j)$ for all $i, j = 1, \dots, N$.

In Equation 3, ω_t is the weight coefficient for every neighboring point. In order to be able to tolerate boundary deformations, the costs of mapping points closer to the two points $\mathbf{p}_i, \mathbf{q}_j$ are treated as more important than the costs of mapping points farther away from $\mathbf{p}_i, \mathbf{q}_j$. To achieve this, the value of ω_t should become smaller and smaller (approaching zero) when t is increasing. In our implementations, the weights of neighboring points are set to decline exponentially, i.e., the weight coefficients are defined as follows:

$$\omega_t = \frac{1}{2^t}. \quad (4)$$

According to Equation 3, we calculate the novel mapping cost for every pair of sample points $\mathbf{p}_i, \mathbf{q}_j$ ($i, j = 1, \dots, N$). With these costs, we obtain a cost matrix $\mathbf{D}(\mathbf{X}, \mathbf{Y})$ for the two shapes \mathbf{X}, \mathbf{Y} :

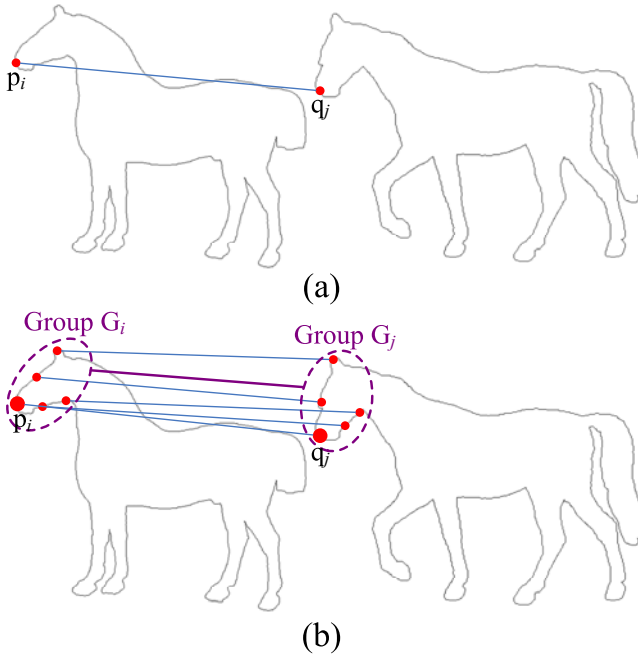


Fig. 2. An example of (a) pair-wised matching and (b) group-wised matching. Note that in our group-wised matching approach the points follow a perfect one-to-one mapping according to the order of the sample points along the shape contours.

$$D(\mathbf{X}, \mathbf{Y}) = (d_{i,j}) = (d(\mathbf{p}_i, \mathbf{q}_j)), i, j = 1, \dots, N. \tag{5}$$

$D(\mathbf{X}, \mathbf{Y})$ is the input for any shape matching algorithm. In our method, we choose the Dynamic Programming (DP) algorithm as the matching method.

2.3 Analysis

Here we give a brief analysis for our method. As introduced in previous sections, the contribution of our method is a more robust matching approach, which results from including more local shape characteristics and the co-occurrence (order) information of contour points into the matching stage. As we see, the original shape descriptor is undefined, thus the original feature distance ($c(\mathbf{p}_i, \mathbf{q}_j)$) is also undefined; this implies that there is no constraint to the original shape descriptor for sample points or local parts. This makes our method very easy to be combined with many current shape descriptors.

There is another valuable property for our method. According to Equation 3, our method can also be regarded as a novel descriptor, which is simply the combination (sequence) of several predefined descriptors. That is, for every sample point \mathbf{p}_i ($i = 1, \dots, N$), we introduce a novel descriptor H_i :

$$H_i = (F_{i-k}, F_{i-k+1}, \dots, F_{i-1}, F_i, F_{i+1}, \dots, F_{i+k-1}, F_{i+k}), \quad (6)$$

where F_t ($t = i-k, \dots, i+k$) is the predefined descriptor for the t th sample point within the group G_i . Then the group-wised mapping cost defined in Equation 3 is just a novel pair-wised mapping cost as the feature distance between the two sample points $\mathbf{p}_i, \mathbf{q}_j$ based on their novel descriptors H_i, H_j :

$$d(\mathbf{p}_i, \mathbf{q}_j) = c(H_i, H_j) = \sum_{t=-k}^k c(F_{i+t}, F_{j+t}). \quad (7)$$

In this way, our method is able to inherit the properties of the primary descriptors. For example, if we adopt the primary descriptor as Shape Context, then the group-wised manner automatically obtains all the properties of Shape Context. This property makes our method very convenient when facing different applications as long as we are able to change the predefined descriptors freely.

The computational complexity of our method remains unchanged compared with the predefined descriptors. This results from Equation 3, which implies that in our method the only work we need to do is to sum up some original costs (feature distance) together to create some new costs. Since for every sample point, the number of costs being added together is the same (which is determined by the parameter k), this process can be implemented by matrix operations, whose computational complexity is even much lower than the complexity of feature distance computation. As a result, our group-wised algorithm doesn't bring any additional computational burden to the whole shape matching system, although it will make some more computational demands.

3 Experiments

This section gives the experimental results using our method. Without losing the generality of our method, the predefined descriptor is chosen as Shape Context (SC) [3], one of the most classical descriptors in the literature. The most important parameter in our method is the group radius k . In our implementations, k is equal to 2 for all situations. We test the effect of the proposed method with three widely used shape databases, i.e., MPEG-7 data set [1], Tari1000 data set [13] and Kimia's 99 data set [5].

3.1 MPEG-7 Shape Database

The database of MPEG-7 CE-Shape-1 part B [1] is very famous in shape matching and classification. This database consists of 1,400 binary images from 70 shape categories, i.e., 20 images per category. This data set is rather difficult since there are some large intra-class variances. Some examples of this data set are given in Figure 3. Following the common performance measurement "bullseye test" [6][9][11][3], we treat every image in this database as a query, and count the number of correct images in the top 40 matches.

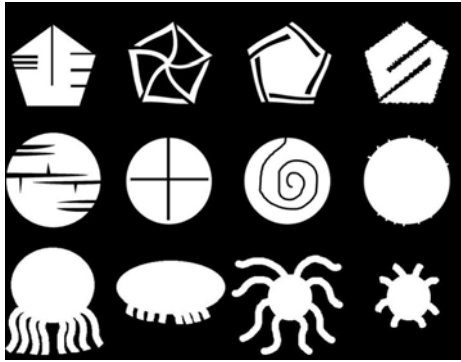


Fig. 3. Example shapes in MPEG-7 CE-Shape-1 part B database [1]

Table 1 presents the retrieval rates of our method and some classical descriptors. We see that the proposed method outperforms Shape Context when our method treats SC as the predefined descriptor. Besides, it outperforms some important works such as Inner-distance Shape Context (IDSC) [9] and Multi-scale Convexity Concavity (MCC) [6] as well. Our method also achieves a considerable retrieval rate compared with some recent works such as Triangle Area Representation (TAR) [11] and Shape Tree [10].

Table 1. Retrieval rates (Bullseye) of different methods on MPEG-7 CE-Shape-1 Part B data set [1]

Algorithm	Score
Shape Tree [10]	87.70%
SC+Group-wised Matching+DP (Ours)	87.15%
TAR [11]	87.13%
SC+DP [15]	86.80%
IDSC+DP [9]	85.40%
MCC [6]	84.93%
Generative Models [7]	80.03%
SC+TPS [3]	76.51%
Visual Parts [2]	76.45%

3.2 Kimia’s 99 Database

The Kimia’s 99 shape data set [5] is also very famous in shape matching and recognition. There are some occlusions, articulations and local deformations, thus it is suitable to be used to check both contour-based and skeleton-based shape descriptors. This data set includes ninety nine binary images from nine categories (see in Fig. 4). The retrieval result is summarized by counting the correct number of top 1 to top 10 nearest matches. The best possible result for each of them is 99.

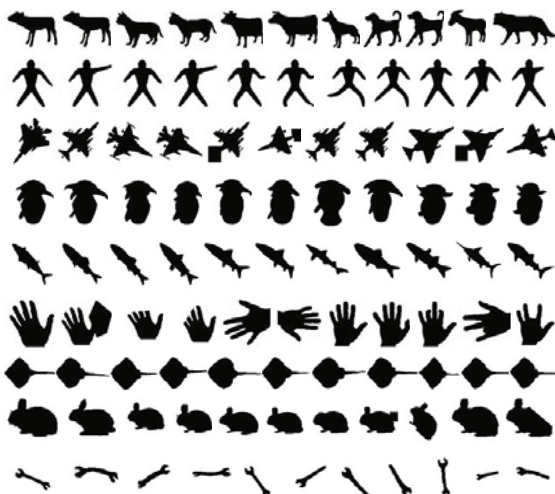


Fig. 4. The ninety-nine shapes in Kimia's 99 database [5]

Table 2. Retrieval results on Kimia's 99 data set [5]

Algorithm	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
SC+TPS [3]	97	91	88	85	84	77	75	66	56	37
Generative Model [7]	99	97	99	98	96	96	94	83	75	48
IDSC+DP [9]	99	99	99	98	98	97	97	98	94	79
SC+DP	98	97	98	98	97	97	96	97	96	80
TAR [8]	99	99	99	98	98	97	98	95	93	80
SC+Group-wised Matching+DP (Ours)	99	97	97	97	96	98	96	96	95	89
Shape Tree [10]	99	99	99	99	99	99	99	97	93	86

Table 2 gives the results of our method and some recent descriptors. Again, we find that our method outperforms Shape Context, and it is comparable with IDSC and TAR. The best result is achieved by Shape Tree [10].

3.3 Tari1000 Database

The Tari1000 data set [13] is also a large database for binary images. It consists of 1,000 binary images from 50 shape categories, i.e., 20 images per category, which is the same as in MPEG-7 database. Some of these categories are also included in MPEG-7 database, such as *brick*, *cattle*, *cellular*, *phone*, *face*, *flatfish*, *fountain*, *key*, *ray*, *teddy*, *watch* and so on. Figure 5 gives some examples for this data set. When testing the retrieval performance, we follow two rules, one is the “bulls-eye test” introduced in Section 3.1, and the other is the Precision-Recall (P-R) curve.

When following the “bulls-eye test”, the score of our method (94.58%) outperforms the result of Shape Context (94.18% [15]). Figure 6 shows the Precision-Recall curves of our method, Shape Context along with some other methods.



Fig. 5. Example shapes in Tari1000 database [13]

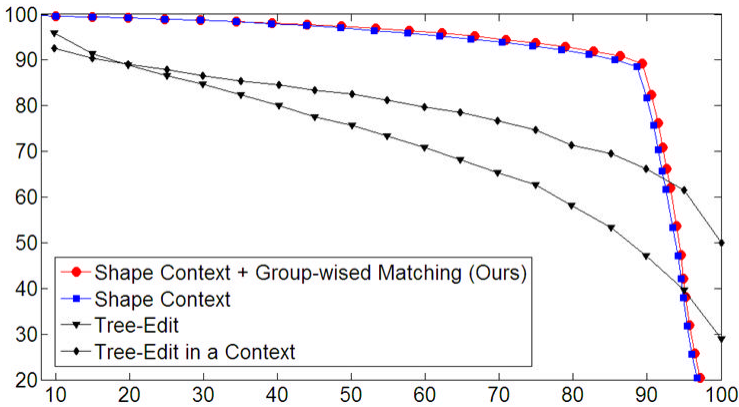


Fig. 6. Precision/Recall curves on Tari1000 database [13]

From the P-R curves we see that our method slightly outperforms Shape Context, and significantly outperforms the methods in [13].

4 Conclusions

This paper presents one novel shape matching method called Group-wised Matching. As shown by its name, this method extends the usual pair-wised matching algorithm to a group-wised matching framework. One point is matched to another when the two corresponding groups achieves an excellent mapping. As the group of one contour sample point consists of its neighboring points, more shape characteristics along with the co-occurrence (order) information of the point sequence on the shape contour are included within the matching stage. As a result, the proposed matching method is able to obtain a more robust matching result. The robustness and power of our method has been demonstrated

by the shape matching and retrieval experiments on three famous benchmarks: MPEG-7 data set, Kimia's data set and Tari1000 data set. The retrieval results show that the new group-wised matching method is able to get encouraging improvements compared to some traditional pair-wised matching approaches.

Several extensions of the proposed approach are possible. In this paper, the mapping cost for two groups is defined in a simple way, as every group is regarded as a short point sequence. In fact, the group may also be considered as a graph, and some current graph matching algorithms can be used to define the group-wised matching cost. Currently, the neighborhood or the size of the group (controlled by the parameter k) for every sample point are completely the same. This strategy, although very simple and easy to implement, may be improved. An algorithm may be invented to calculate one suitable value for the parameter k for each sample point. Next, the group-wised manner may not only be used for contour points, but also be used for shape instances. Finally, our method may be extended to some other problems and applications such as point pattern matching [4] [12] and object detection in real images [16] [14].

Acknowledgments. The authors would like to thank Ling and Jacobs for releasing their source code of IDSC online. Parts of our work (such as boundary extraction, calculation of the Shape Context and Dynamic Programming algorithm) are accomplished by directly using corresponding parts of this code or slightly modifying them. The authors also would like to thank the anonymous referees who gave us many helpful comments and suggestions. This work was supported by National Natural Science Foundation of China #60903096 and #60873127.

References

1. Latecki, L.J., Lakamper, R., Eckhardt, U.: Shape Descriptors for Non-rigid Shapes with A Single Closed Contour. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 424–429 (2000)
2. Latecki, L.J., Lakamper, R.: Shape Similarity Measure Based on Correspondence of Visual Parts. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 22, 1185–1190 (2000)
3. Belongie, S., Malik, J., Puzicha, J.: Shape Matching and Object Recognition Using Shape Context. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 24, 509–522 (2002)
4. Chui, H., Rangarajan, A.: A New Point Matching Algorithm for Non-rigid Registration. Computer Vision and Image Understanding (CVIU) 89(2), 114–141 (2003)
5. Sebastian, T.B., Klein, P.N., Kimia, B.B.: Recognition of Shapes by Editing Their Shock Graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 25, 116–125 (2004)
6. Adamek, T., O'Connor, N.E.: A Multiscale Representation Method for Nonrigid Shapes with a Single Closed Contour. IEEE Transactions on Circuits and Systems for Video Technology (TCSVT) 14, 742–753 (2004)
7. Tu, Z., Yuille, A.L.: Shape Matching and Recognition – Using Generative Models and Informative Features. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004, Part III. LNCS, vol. 3023, pp. 195–209. Springer, Heidelberg (2004)

8. Alajlan, N., Rube, I.E., Kamel, M.S., Freeman, G.: Shape Retrieval Using Triangle-area Representation and Dynamic Space Warping. *Pattern Recognition (PR)* 40, 1911–1920 (2007)
9. Ling, H., Jacobs, D.W.: Shape Classification Using Inner-Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 29, 286–299 (2007)
10. Felzenszwalb, P.F., Schwartz, J.D.: Hierarchical Matching of Deformable Shapes. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–8 (2007)
11. Alajlan, N., Kamel, M., Freeman, G.: Geometry-based Image Retrieval in Binary Image Databases. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 30, 1003–1013 (2008)
12. Mcauley, J.J., Caetano, T.S., Barbosa, M.S.: Graph Rigidity, Cyclic Belief Propagation, and Point Pattern Matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 30(11), 2047–2054 (2008)
13. Baseski, E., Erdem, A., Tari, S.: Dissimilarity between Two Skeletal Trees in A Context. *Pattern Recognition (PR)* 42, 370–385 (2009)
14. Bai, X., Wang, X., Latecki, L.J., Liu, W., Tu, Z.: Active Skeleton for Non-rigid Object Detection. In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 575–582 (2009)
15. Bai, X., Wang, B., Wang, X., Liu, W., Tu, Z.: Co-transduction for shape retrieval. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010, Part III. LNCS*, vol. 6313, pp. 328–341. Springer, Heidelberg (2010)
16. Wang, B., Bai, X., Wang, X., Liu, W., Tu, Z.: Object Recognition Using Junctions. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010, Part V. LNCS*, vol. 6315, pp. 15–28. Springer, Heidelberg (2010)

Author Index

- Adluru, Nagesh I-236
Alexander, Andrew L. I-236
Aoki, Naokazu II-180
Aranda, Ramón II-36
Ariki, Yasuo I-97
Asada, Naoki II-323
Atkinson, Gary I-109
Au, Oscar C. II-48
- Bai, Xiang II-393
Bleyer, Michael I-165
Börner, Anko I-224
Boufama, Boubakeur II-168
- Carrasco, Miguel II-192
Castanheira de Souza, Rafael Henrique
I-311
Chang, I-Cheng II-128
Chang, Pao-Chi II-252
Chen, Chun-Hung I-144
Chen, Hsiao-Wei II-335
Chen, Ke-Yin II-128
Chen, Wei-Chih II-381
Chen, Yung-Chang I-48
Chew, S.W. II-311
Chiang, Jui-Chiu I-144
Chiao, Yen-Hao II-240
Choi, Byeongho II-1
Chung, Moo K. I-36, I-236
- Dai, Wei II-48
Daribo, Ismael II-323
Davidson, Richard J. I-36
de Groen, Piet C. I-61
DinhQuoc, Khanh II-347
Du, Weiwei II-149
Duchaineau, Mark A. I-153
- Elamsy, Tarik II-168
- Fang, Jiunn-Tsair II-252
Fujiyoshi, Masaaki II-180
Fukui, Kazuhiro II-82
Furukawa, Ryo II-323
- Gelautz, Margrit I-165
Geng, Haokun II-274
Ghimire, Deepak 1
Gil, Jong In I-13
- Ha, Synh Viet Uyen I-323, II-60
Habel, Adlane II-168
Hayet, Jean-Bernard II-287
He, Xiangjian I-214
Heng, Swee-Huay I-257
Heo, Jin I-267
Hermann, Simon I-224, I-395
Hess-Flores, Mauricio I-153
Hiraki, Kazuo I-277
Hiura, Shinsaku II-323
Ho, Yo-Sung I-121, I-267, I-301, I-384,
II-25
HoangVan, Xiem II-347
Hong, Min-Cheol II-157
Hosni, Asmaa I-165
Hsu, Shih-Chung II-116, II-128
Huang, Chung-Lin II-116, II-128
Huang, Po-Hao II-240
Huang, Sheng-Yen II-381
- Imaizumi, Shoko II-180
Iwai, Yoshio II-262
- Jamzad, Mansour II-359
Jang, Seung Eun I-13
Jawed, Khurram I-202
Jeon, Byeungwoo II-347
Jeon, Jae Wook I-73, I-85, I-323, II-60
Jeong, Jechang II-1
Jia, Weijia II-92
Jia, Wenjing I-214
Jiang, Gangyi II-227
Jiang, Yan-Ting II-252
Joy, Kenneth I. I-153
- Kamata, Sei-ichiro II-141
Kang, Yousun I-248
Kang, Yun-Suk I-301
Kawasaki, Hiroshi II-323
Kim, Je-Woo II-1, II-13
Kim, Manbae I-13

- Kim, Seung-Goo I-36
 Kim, Yong-Hwan II-13
 Kimura, Ryohei II-262
 Kiya, Hitoshi II-180
 Klette, Reinhard I-224, I-395, II-274
 Knoblauch, Daniel I-153
 Kobayashi, Hiroyuki II-180
 Komai, Yuto I-97
 Kuester, Falko I-153

 Lai, Shang-Hong II-215, II-240, II-299,
 II-335
 Lainhart, Janet E. I-236
 Lange, Nicholas T. I-236
 Lee, Gyo-Yoon I-384
 Lee, Jaejoon II-371
 Lee, Joonwhoan I-1
 Leitão, Helena C.G. I-109
 Li, Fucui II-227
 Li, Guiqing II-92
 Li, Ke-Chun II-215
 Li, Sijin II-48
 Lie, Wen-Nung I-144, II-381
 Lin, Guo-Shiang II-381
 Lin, Shang-Yen I-48
 Lin, Shen-Ju II-116
 Ling, Huo-Chong I-257
 Liu, Han-Wen II-381
 Liu, Wenyu II-393
 Lucey, P. II-311
 Lucey, S. II-311

 Madrigal, Francisco II-287
 Matsui, Shuhei I-335
 May, Michael I-190, I-289
 Mery, Domingo II-192
 Migita, Tsuyoshi I-178
 Mori, Shiya II-149
 Morris, John I-202
 Morris, Tim I-190, I-289
 Muthukudage, Jayantha I-61

 Naemura, Takeshi I-22, I-407
 Nagahara, Hajime I-335
 Naito, Takashi I-248
 Nakamori, Nobuyuki II-149
 Nakashima, Ryo I-407
 Nguyen, Duc Dung I-73
 Nguyen, Thuy Tuong I-85
 Nguyen, Tuan-Anh II-157

 Nicolescu, Radu II-274
 Ninomiya, Yoshiki I-248
 Nishihara, Akinori II-71
 Nosaka, Ryusuke II-82

 Oh, JungHwan I-61
 Oh, Kwan-Jung II-371
 Ohkawa, Yasuhiro II-82
 Okabe, Takahiro I-277
 Okutomi, Masatoshi I-311

 Park, Du-Sik II-371
 Park, Jiho II-1, II-13
 Peng, Zongju II-227
 Pham, Cuong Cao I-323, II-60
 Phan, Raphael C.-W. I-257

 Ramírez-Manzanares, Alonso II-36
 Rana, R. II-311
 Rhemann, Christoph I-165
 Rivera, Mariano II-36, II-287
 Russell, James II-274

 Sagawa, Ryusuke II-323
 Sakazawa, Shigeyuki I-370
 Saracchini, Rafael F.V. I-109
 Sato, Kosuke II-262
 Sato, Yoichi I-277
 Schaefer, Stacey M. I-36
 Seo, Seongho I-36
 Seto, Hiroyuki II-104
 Shakeri, Mahsa II-359
 Shakunaga, Takeshi I-178, II-104
 Shao, Feng II-227
 Shin, Bok-Suk II-274
 Shin, In-Yong I-121
 Sim, Jae-Young II-204
 Smith, Melvyn L. I-109
 Sogawa, Kazuhiro I-178
 Sridharan, S. II-311
 Srinark, Thitiwan I-358
 Sthitpattanapongsa, Puthipong I-358
 Stolfi, Jorge I-109
 Su, Hong-Ren II-215, II-299
 Sugano, Yusuke I-277
 Sugimoto, Akihiro I-277
 Sun, Lin II-48

 Tae-o-sot, Sarawut II-71
 Taguchi, Tomoyuki II-104

- Tai, Jason I-348
Takagi, Koichi I-370
Takahashi, Keita I-22, I-407
Takemura, Noriko II-262
Takiguchi, Tetsuya I-97
Taniguchi, Rin-ichiro I-335
Tavanapong, Wallapak I-61
Torii, Akihiko I-311
Turner, Martin I-190, I-289

van Reekum, Carien M. I-36

Wang, Dong II-92
Wang, Junwei II-393
Wang, Ren-Jie II-252
Wang, Sheng I-214
Wong, Johnny I-61
Wu, Qiang I-214

Xiong, Yunhui II-92
Xu, Jianfeng I-370

Yamada, Kentaro I-277
Yamaguchi, Koichiro I-248
Yang, Hao-Liang II-240
Yang, Zhuo II-141
Yao, Anbang I-132
Yeh, Mei-Chen I-348
Yoon, Da-Hyun II-25
Yu, Mei II-227
Yu, Shan I-132

Zheng, Qiaoyan II-227
Zhou, Yu II-393
Zhu, Jiangying II-227
Zou, Ruobing II-48