

LNCS 6523

Kuo-Tien Lee Wen-Hsiang Tsai
Hong-Yuan Mark Liao Tsuhan Chen
Jun-Wei Hsieh Chien-Cheng Tseng (Eds.)

Advances in Multimedia Modeling

17th International Multimedia Modeling Conference, MMM 2011
Taipei, Taiwan, January 2011
Proceedings, Part I

1
Part I



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

Kuo-Tien Lee Wen-Hsiang Tsai
Hong-Yuan Mark Liao Tsuhan Chen
Jun-Wei Hsieh Chien-Cheng Tseng (Eds.)

Advances in Multimedia Modeling

17th International
Multimedia Modeling Conference, MMM 2011
Taipei, Taiwan, January 5-7, 2011
Proceedings, Part I



Springer

Volume Editors

Kuo-Tien Lee
Jun-Wei Hsieh
National Taiwan Ocean University
Keelung, Taiwan
E-mail: {po,shieh}@mail.ntou.edu.tw

Wen-Hsiang Tsai
National Chiao Tung University
Hsinchu, Taiwan
E-mail: whtsai@cis.nctu.edu.tw

Hong-Yuan Mark Liao
Academia Sinica
Taipei, Taiwan
E-mail: liao@iis.sinica.edu.tw

Tsuhan Chen
Cornell University
Ithaca, NY, USA
E-mail: tc234@cornell.edu

Chien-Cheng Tseng
National Kaohsiung First University of Science and Technology
Kaohsiung, Taiwan
E-mail: tcc@ccms.nkfust.edu.tw

Library of Congress Control Number: 2010940989

CR Subject Classification (1998): H.5.1, I.5, H.3, H.4, I.4, H.2.8

LNCS Sublibrary: SL 3 – Information Systems and Application, incl. Internet/Web and HCI

ISSN 0302-9743
ISBN-10 3-642-17831-6 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-17831-3 Springer Berlin Heidelberg New York

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.

springer.com

© Springer-Verlag Berlin Heidelberg 2011
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper 06/3180

Preface

Welcome to the proceedings of the 17th Multimedia Modeling Conference (MMM 2011) held in Taipei, Taiwan, during January 5–7, 2011. Following the success of the 16 preceding conferences, the 17th MMM brought together researchers, developers, practitioners, and educators in the field of multimedia. Both practical systems and theories were presented in this conference, thanks to the support of Microsoft Research Asia, Industrial Technology Research Institute, Institute for Information Industry, National Museum of Natural Science, and the Image Processing and Pattern Recognition Society of Taiwan.

MMM 2011 featured a comprehensive program including keynote speeches, regular paper presentations, posters, and special sessions. We received 450 papers in total. Among these submissions, we accepted 75 oral presentations and 21 poster presentations. Six special sessions were organized by world-leading researchers. We sincerely acknowledge the Program Committee members who have contributed much of their precious time during the paper reviewing process.

We would like to sincerely thank the support of our strong Organizing Committee and Advisory Committee. Special thanks go to Jun-Wei Hsieh, Tun-Wen Pai, Shyi-Chyi Cheng, Hui-Huang Hsu, Tao Mei, Meng Wang, Chih-Min Chao, Chun-Chao Yeh, Shu-Hsin Liao, and Chin-Chun Chang. This conference would never have happened without their help.

January 2011

Wen-Hsiang Tsai
Mark Liao
Kuo-Tien Lee

Organization

MMM 2011 was hosted and organized by the Department of Computer Science and Engineering, National Taiwan Ocean University, Taiwan. The conference was held at the National Taiwan Science Education Center, Taipei, during January 5–7, 2011.

Conference Committee

Steering Committee	Yi-Ping Phoebe Chen (La Trobe University, Australia) Tat-Seng Chua (National University of Singapore, Singapore) Tosiyasu L. Kunii (University of Tokyo, Japan) Wei-Ying Ma (Microsoft Research Asia, China) Nadia Magnenat-Thalmann (University of Geneva, Switzerland)
Conference Co-chairs	Patrick Senac (ENSICA, France) Kuo-Tien Lee (National Taiwan Ocean University, Taiwan) Wen-Hsiang Tsai (National Chiao Tung University, Taiwan)
Program Co-chairs	Hong-Yuan Mark Liao (Academia Sinica, Taiwan) Tsuhan Chen (Cornell University, USA) Jun-Wei Hsieh (National Taiwan Ocean University, Taiwan)
Special Session Co-chairs	Chien-Cheng Tseng (National Kaohsiung First University of Science and Technology, Taiwan) Hui-Huang Hsu (Tamkang University, Taiwan) Tao Mei (Microsoft Research Asia, China)
Demo Co-chair	Meng Wang (Microsoft Research Asia, China)
Local Organizing Co-chairs	Tun-Wen Pai (National Taiwan Ocean University, Taiwan) Shyi-Chyi Cheng (National Taiwan Ocean University, Taiwan)
Publication Chair	Chih-Min Chao (National Taiwan Ocean University, Taiwan)
Publicity Chair	Shu-Hsin Liao (National Taiwan Ocean University, Taiwan)

VIII Organization

US Liaison	Qi Tian (University of Texas at San Antonio, USA)
Asian Liaison	Tat-Seng Chua (National University of Singapore, Singapore)
European Liaison	Susanne Boll (University of Oldenburg, Germany)
Webmaster	Chun-Chao Yeh (National Taiwan Ocean University, Taiwan)

Program Committee

Allan Hanbury	Vienna University of Technology, Austria
Andreas Henrich	University of Bamberg, Germany
Bernard Merialdo	EURECOM, France
Brigitte Kerhervé	University of Quebec, Canada
Cathal Gurrin	Dublin City University, Ireland
Cees Snoek	University of Amsterdam, The Netherlands
Cha Zhang	Microsoft Research
Chabane Djeraba	University of Sciences and Technologies of Lille, France
Changhu Wang	University of Science and Technology of China
Changsheng Xu	NLPR, Chinese Academy of Science, China
Chia-Wen Lin	National Tsing Hua University, Taiwan
Chong-Wah Ngo	City University of Hong Kong, Hong Kong
Christian Timmerer	University of Klagenfurt, Austria
Colum Foley	Dublin City University, Ireland
Daniel Thalmann	EPFL, Swiss
David Vallet	Universidad Autónoma de Madrid, Spain
Duy-Dinh Le	National Institute of Informatics, Japan
Fernando Pereira	Technical University of Lisbon, Portugal
Francisco Jose Silva Mata	Centro de Aplicaciones de Tecnologias de Avanzada, Cuba
Georg Thallinger	Joanneum Research, Austria
Guntur Ravindra	Applied Research & Technology Center, Motorola, Bangalore
Guo-Jun Qi	University of Science and Technology of China
Harald Kosch	Passau University, Germany
Hui-Huang Hsu	Tamkang University, Taiwan
Jen-Chin Jiang	National Dong Hwa University, Taiwan
Jia-hung Ye	National Sun Yat-sen University, Taiwan
Jianmin Li	Tsinghua University, China
Jianping Fan	University of North Carolina, USA
Jiebo Luo	Kodak Research, USA
Jing-Ming Guo	National Taiwan University of Science and Technology, Taiwan
Jinhui Tang	University of Science and Technology of China

Jinjun Wang	NEC Laboratories America, Inc., USA
Jiro Katto	Waseda University, Japan
Joemon Jose	University of Glasgow, UK
Jonathon Hare	University of Southampton, UK
Joo Hwee Lim	Institute for Infocomm Research, Singapore
Jose Martinez	UAM, Spain
Keiji Yanai	University of Electro-Communications, Japan
Koichi Shinoda	Tokyo Institute of Technology, Japan
Lap-Pui Chau	Nanyang Technological University, Singapore
Laura Hollink	Vrije Universiteit Amsterdam, The Netherlands
Laurent Amsaleg	CNRS-IRISA, France
Lekha Chaisorn	Institute for Infocomm Research, Singapore
Liang-Tien Chia	Nanyang Technological University, Singapore
Marcel Worring	University of Amsterdam, The Netherlands
Marco Bertini	University of Florence, Italy
Marco Paleari	EURECOM, France
Markus Koskela	Helsinki University of Technology, Finland
Masashi Inoue	Yamagata University, Japan
Matthew Cooper	FX Palo Alto Lab, Inc., Germany
Matthias Rauterberg	Technical University Eindhoven, The Netherlands
Michael Lew	Leiden University, The Netherlands
Michel Crucianu	Conservatoire National des Arts et Métiers, France
Michel Kieffer	Laboratoire des Signaux et Systèmes, CNRS-Supélec, France
Ming-Huei Jin	Institute for Information Industry, Taiwan
Mohan Kankanhalli	National University of Singapore
Neil O'Hare	Dublin City University, Ireland
Nicholas Evans	EURECOM, France
Noel O'Connor	Dublin City University, Ireland
Nouha Bouteldja	Conservatoire National des Arts et Métiers, France
Ola Stockfelt	Gothenburg University, Sweden
Paul Ferguson	Dublin City University, Ireland
Qi Tian	University of Texas at San Antonio, USA
Raphael Troncy	CWI, The Netherlands
Roger Zimmermann	University of Southern California, USA
Selim Balci soy	Sabanci University, Turkey
Sengamedu Srinivasan	Yahoo! India
Seon Ho Kim	University of Denver, USA
Shen-wen Shr	National Chi Nan University, Taiwan
Shingo Uchihashi	Fuji Xerox Co., Ltd., Japan
Shin'ichi Satoh	National Institute of Informatics, Japan

X Organization

Shiuan-Ting Jang	National Yunlin University of Science and Technology, Taiwan
Shuicheng Yan	National University of Singapore
Shu-Yuan Chen	Yuan Ze University, Taiwan
Sid-Ahmed Berrani	Orange Labs - France Telecom
Stefano Bocconi	Università degli studi di Torino, Italy
Susu Yao	Institute for Infocomm Research, Singapore
Suzanne Little	Open University, UK
Tao Mei	Microsoft Research Asia, China
Taro Tezuka	Ritsumeikan University, Japan
Tat-Seng Chua	National University of Singapore
Thierry Pun	University of Geneva, Switzerland
Tong Zhang	HP Labs
Valerie Gouet-Brunet	Conservatoire National des Arts et Metiers, France
Vincent Charvillat	University of Toulouse, France
Vincent Oria	NJIT, USA
Wai-tian Tan	Hewlett-Packard, USA
Wei Cheng	University of Michigan, USA
Weiqi Yan	Queen's University Belfast, UK
Weisi Lin	Nanyang Technological University, Singapore
Wen-Hung Liau	National Chengchi University, Taiwan
Werner Bailer	Joanneum Research, Austria
William Grosky	University of Michigan, USA
Winston Hsu	National Taiwan University, Taiwan
Wolfgang Hürst	Utrecht University, The Netherlands
Xin-Jing Wang	Microsoft Research Asia, China
Yannick Prié	LIRIS, France
Yan-Tao Zheng	National University of Singapore, Singapore
Yea-Shuan Huang	Chung-Hua University, Taiwan
Yiannis Kompatsiaris	Informatics and Telematics Institute Centre for Research and Technology Hellas, Greece
Yijuan Lu	Texas State University, USA
Yongwei Zhu	Institute for Infocomm Research Asia, Singapore
Yun Fu	University at Buffalo (SUNY), USA
Zha Zhengjun	National University of Singapore, Singapore
Zheng-Jun Zha	National University of Singapore, Singapore
Zhongfei Zhang	State University of New York at Binghamton, USA
Zhu Li	Hong Kong Polytechnic University, Hong Kong

Sponsors

Microsoft Research
Industrial Technology Research Institute
Institute For Information Industry
National Taiwan Science Education Center
National Taiwan Ocean University
Bureau of Foreign Trade
National Science Council

Table of Contents – Part I

Regular Papers

Audio, Image, Video Processing, Coding and Compression

A Generalized Coding Artifacts and Noise Removal Algorithm for Digitally Compressed Video Signals	1
<i>Ling Shao, Hui Zhang, and Yan Liu</i>	
Efficient Mode Selection with BMA Based Pre-processing Algorithms for H.264/AVC Fast Intra Mode Decision	10
<i>Chen-Hsien Miao and Chih-Peng Fan</i>	
Perceptual Motivated Coding Strategy for Quality Consistency	21
<i>Like Yu, Feng Dai, Yongdong Zhang, and Shouxun Lin</i>	
Compressed-Domain Shot Boundary Detection for H.264/AVC Using Intra Partitioning Maps	29
<i>Sarah De Bruyne, Jan De Cock, Chris Poppe, Charles-Frederik Hollemeersch, Peter Lambert, and Rik Van de Walle</i>	
Adaptive Orthogonal Transform for Motion Compensation Residual in Video Compression	40
<i>Zhouye Gu, Weisi Lin, Bu-sung Lee, and Chiew Tong Lau</i>	
Parallel Deblocking Filter for H.264/AVC on the TILERa Many-Core Systems	51
<i>Chenggang Yan, Feng Dai, and Yongdong Zhang</i>	
Image Distortion Estimation by Hash Comparison	62
<i>Li Weng and Bart Preneel</i>	

Media Content Browsing and Retrieval

Sewing Photos: Smooth Transition between Photos	73
<i>Tzu-Hao Kuo, Chun-Yu Tsai, Kai-Yin Cheng, and Bing-Yu Chen</i>	
Employing Aesthetic Principles for Automatic Photo Book Layout	84
<i>Philipp Sandhaus, Mohammad Rabbath, and Susanne Boll</i>	
Video Event Retrieval from a Small Number of Examples Using Rough Set Theory	96
<i>Kimiaki Shirahama, Yuta Matsuoka, and Kuniaki Uehara</i>	

Community Discovery from Movie and Its Application to Poster Generation	107
<i>Yan Wang, Tao Mei, and Xian-Sheng Hua</i>	
A BOVW Based Query Generative Model	118
<i>Reede Ren, John Collomosse, and Joemon Jose</i>	
Video Sequence Identification in TV Broadcasts	129
<i>Klaus Schoeffmann and Laszlo Boeszoermenyi</i>	
Content-Based Multimedia Retrieval in the Presence of Unknown User Preferences	140
<i>Christian Beecks, Ira Assent, and Thomas Seidl</i>	
Multi-Camera, Multi-View, and 3D Systems	
People Localization in a Camera Network Combining Background Subtraction and Scene-Aware Human Detection	151
<i>Tung-Ying Lee, Tsung-Yu Lin, Szu-Hao Huang, Shang-Hong Lai, and Shang-Chih Hung</i>	
A Novel Depth-Image Based View Synthesis Scheme for Multiview and 3DTV	161
<i>Xun He, Xin Jin, Minghui Wang, and Satoshi Goto</i>	
Egocentric View Transition for Video Monitoring in a Distributed Camera Network	171
<i>Kuan-Wen Chen, Pei-Jyun Lee, and Yi-Ping Hung</i>	
A Multiple Camera System with Real-Time Volume Reconstruction for Articulated Skeleton Pose Tracking	182
<i>Zheng Zhang, Hock Soon Seah, Chee Kwang Quah, Alex Ong, and Khalid Jabbar</i>	
A New Two-Omni-Camera System with a Console Table for Versatile 3D Vision Applications and Its Automatic Adaptation to Imprecise Camera Setups	193
<i>Shen-En Shih and Wen-Hsiang Tsai</i>	
3D Face Recognition Based on Local Shape Patterns and Sparse Representation Classifier	206
<i>Di Huang, Karima Ouji, Mohsen Ardabilian, Yunhong Wang, and Liming Chen</i>	
An Effective Approach to Pose Invariant 3D Face Recognition	217
<i>Dayong Wang, Steven C.H. Hoi, and Ying He</i>	

Multimedia Indexing and Mining

Score Following and Retrieval Based on Chroma and Octave Representation	229
<i>Wei-Ta Chu and Meng-Luen Li</i>	
Incremental Multiple Classifier Active Learning for Concept Indexing in Images and Videos	240
<i>Bahjat Safadi, Yubing Tong, and Georges Quénot</i>	
A Semantic Higher-Level Visual Representation for Object Recognition	251
<i>Ismail El Sayad, Jean Martinet, Thierry Urruty, and Chabane Dejraha</i>	
Mining Travel Patterns from GPS-Tagged Photos	262
<i>Yan-Tao Zheng, Yiqun Li, Zheng-Jun Zha, and Tat-Seng Chua</i>	
Augmenting Image Processing with Social Tag Mining for Landmark Recognition	273
<i>Amogh Mahapatra, Xin Wan, Yonghong Tian, and Jaideep Srivastava</i>	
News Shot Cloud: Ranking TV News Shots by Cross TV-Channel Filtering for Efficient Browsing of Large-Scale News Video Archives	284
<i>Norio Katayama, Hiroshi Mo, and Shin'ichi Satoh</i>	

Multimedia Content Analysis (I)

Speaker Change Detection Using Variable Segments for Video Indexing	296
<i>King Yiu Tam, Jose Lay, and David Levy</i>	
Correlated PLSA for Image Clustering	307
<i>Peng Li, Jian Cheng, Zechao Li, and Hanqing Lu</i>	
Genre Classification and the Invariance of MFCC Features to Key and Tempo	317
<i>Tom L.H. Li and Antoni B. Chan</i>	
Combination of Local and Global Features for Near-Duplicate Detection	328
<i>Yue Wang, ZuJun Hou, Kariantto Leman, Nam Trung Pham, TeckWee Chua, and Richard Chang</i>	
Audio Tag Annotation and Retrieval Using Tag Count Information	339
<i>Hung-Yi Lo, Shou-De Lin, and Hsin-Min Wang</i>	

Similarity Measurement for Animation Movies	350
<i>Alexandre Benoit, Madalina Ciobotaru, Patrick Lambert, and Bogdan Ionescu</i>	

Multimedia Content Analysis (II)

A Feature Sequence Kernel for Video Concept Classification	359
<i>Werner Bailer</i>	
Bottom-Up Saliency Detection Model Based on Amplitude Spectrum ...	370
<i>Yuming Fang, Weisi Lin, Bu-Sung Lee, Chiew Tong Lau, and Chia-Wen Lin</i>	
L_2 -Signature Quadratic Form Distance for Efficient Query Processing in Very Large Multimedia Databases	381
<i>Christian Beecks, Merih Seran Uysal, and Thomas Seidl</i>	
Generating Representative Views of Landmarks via Scenic Theme Detection	392
<i>Yi-Liang Zhao, Yan-Tao Zheng, Xiangdong Zhou, and Tat-Seng Chua</i>	
Regularized Semi-supervised Latent Dirichlet Allocation for Visual Concept Learning	403
<i>Liansheng Zhuang, Lanbo She, Jingjing Huang, Jiebo Luo, and Nenghai Yu</i>	
Boosted Scene Categorization Approach by Adjusting Inner Structures and Outer Weights of Weak Classifiers	413
<i>Xueming Qian, Zhe Yan, and Kaiyu Hang</i>	
A User-Centric System for Home Movie Summarisation	424
<i>Saman H. Cooray, Hyowon Lee, and Noel E. O'Connor</i>	

Multimedia Signal Processing and Communications

Image Super-Resolution by Vectorizing Edges.....	435
<i>Chia-Jung Hung, Chun-Kai Huang, and Bing-Yu Chen</i>	
Vehicle Counting without Background Modeling	446
<i>Cheng-Chang Lien, Ya-Ting Tsai, Ming-Hsiu Tsai, and Lih-Guong Jang</i>	
Effective Color-Difference-Based Interpolation Algorithm for CFA Image Demosaicking.....	457
<i>Yea-Shuan Huang and Sheng-Yi Cheng</i>	

Utility Max-Min Fair Rate Allocation for Multiuser Multimedia Communications	470
<i>Qing Zhang, Guizhong Liu, and Fan Li</i>	
Multimedia Applications	
Adaptive Model for Robust Pedestrian Counting	481
<i>Jingjing Liu, Jinqiao Wang, and Hanqing Lu</i>	
Multi Objective Optimization Based Fast Motion Detector	492
<i>Jia Su, Xin Wei, Xiaocong Jin, and Takeshi Ikenaga</i>	
Narrative Generation by Repurposing Digital Videos	503
<i>Nick C. Tang, Hsiao-Rong Tyan, Chiou-Ting Hsu, and Hong-Yuan Mark Liao</i>	
A Coordinate Transformation System Based on the Human Feature Information	514
<i>Shih-Ming Chang, Joseph Tsai, Timothy K. Shih, and Hui-Huang Hsu</i>	
An Effective Illumination Compensation Method for Face Recognition	525
<i>Yea-Shuan Huang and Chu-Yung Li</i>	
Shape Stylized Face Caricatures	536
<i>Nguyen Kim Hai Le, Yong Peng Why, and Golam Ashraf</i>	
<i>i-m-Breath</i> : The Effect of Multimedia Biofeedback on Learning Abdominal Breath	548
<i>Meng-Chieh Yu, Jin-Shing Chen, King-Jen Chang, Su-Chu Hsu, Ming-Sui Lee, and Yi-Ping Hung</i>	
Author Index	559

Table of Contents – Part II

Special Session Papers

Content Analysis for Human-Centered Multimedia Applications

Generative Group Activity Analysis with Quaternion Descriptor	1
<i>Guangyu Zhu, Shuicheng Yan, Tony X. Han, and Changsheng Xu</i>	
Grid-Based Retargeting with Transformation Consistency Smoothing ...	12
<i>Bing Li, Ling-Yu Duan, Jinqiao Wang, Jie Chen, Rongrong Ji, and Wen Gao</i>	
Understanding Video Sequences through Super-Resolution	25
<i>Yu Peng, Jesse S. Jin, Suhuai Luo, and Mira Park</i>	
Facial Expression Recognition on Hexagonal Structure Using LBP-Based Histogram Variances	35
<i>Lin Wang, Xiangjian He, Ruo Du, Wenjing Jia, Qiang Wu, and Wei-chang Yeh</i>	

Mining Social Relationship from Media Collections

Towards More Precise Social Image-Tag Alignment	46
<i>Ning Zhou, Jinye Peng, Xiaoyi Feng, and Jianping Fan</i>	
Social Community Detection from Photo Collections Using Bayesian Overlapping Subspace Clustering	57
<i>Peng Wu, Qiang Fu, and Feng Tang</i>	
Dynamic Estimation of Family Relations from Photos	65
<i>Tong Zhang, Hui Chao, and Dan Tretter</i>	

Large Scale Rich Media Data Management

Semi-automatic Flickr Group Suggestion	77
<i>Junjie Cai, Zheng-Jun Zha, Qi Tian, and Zengfu Wang</i>	
A Visualized Communication System Using Cross-Media Semantic Association	88
<i>Xinming Zhang, Yang Liu, Chao Liang, and Changsheng Xu</i>	

Effective Large Scale Text Retrieval via Learning Risk-Minimization and Dependency-Embedded Model	99
<i>Sheng Gao and Haizhou Li</i>	

Efficient Large-Scale Image Data Set Exploration: Visual Concept Network and Image Summarization	111
<i>Chunlei Yang, Xiaoyi Feng, Jinye Peng, and Jianping Fan</i>	

Multimedia Understanding for Consumer Electronics

A Study in User-Centered Design and Evaluation of Mental Tasks for BCI	122
<i>Danny Plass-Oude Bos, Mannes Poel, and Anton Nijholt</i>	

Video CooKing: Towards the Synthesis of Multimedia Cooking Recipes	135
<i>Keisuke Doman, Cheng Ying Kuai, Tomokazu Takahashi, Ichiro Ide, and Hiroshi Murase</i>	

Snap2Read: Automatic Magazine Capturing and Analysis for Adaptive Mobile Reading	146
<i>Yu-Ming Hsu, Yen-Liang Lin, Winston H. Hsu, and Brian Wang</i>	

Multimodal Interaction Concepts for Mobile Augmented Reality Applications	157
<i>Wolfgang Hürst and Casper van Wezel</i>	

Image Object Recognition and Compression

Morphology-Based Shape Adaptive Compression	168
<i>Jian-Jiun Ding, Pao-Yen Lin, Jiun-De Huang, Tzu-Heng Lee, and Hsin-Hui Chen</i>	

People Tracking in a Building Using Color Histogram Classifiers and Gaussian Weighted Individual Separation Approaches	177
<i>Che-Hung Lin, Sheng-Luen Chung, and Jing-Ming Guo</i>	

Human-Centred Fingertip Mandarin Input System Using Single Camera	187
<i>Chih-Chang Yu, Hsu-Yung Cheng, Bor-Shenn Jeng, Chien-Cheng Lee, and Wei-Tyng Hong</i>	

Automatic Container Code Recognition Using Compressed Sensing Method	196
<i>Chien-Cheng Tseng and Su-Ling Lee</i>	

Combining Histograms of Oriented Gradients with Global Feature for Human Detection	208
<i>Shih-Shinh Huang, Hsin-Ming Tsai, Pei-Yung Hsiao, Meng-Qui Tu, and Er-Liang Jian</i>	
Interactive Image and Video Search	
Video Browsing Using Object Trajectories	219
<i>Felix Lee and Werner Bailer</i>	
Size Matters! How Thumbnail Number, Size, and Motion Influence Mobile Video Retrieval	230
<i>Wolfgang Hürst, Cees G.M. Snoek, Willem-Jan Spoel, and Mate Tomin</i>	
An Information Foraging Theory Based User Study of an Adaptive User Interaction Framework for Content-Based Image Retrieval	241
<i>Haiming Liu, Paul Mulholland, Dawei Song, Victoria Uren, and Stefan Rüger</i>	
Poster Session Papers	
Generalized Zigzag Scanning Algorithm for Non-square Blocks	252
<i>Jian-Jiun Ding, Pao-Yen Lin, and Hsin-Hui Chen</i>	
The Interaction Ontology Model: Supporting the <i>Virtual Director</i> Orchestrating Real-Time Group Interaction	263
<i>Rene Kaiser, Claudia Wagner, Martin Hoeffernig, and Harald Mayer</i>	
CLUENET: Enabling Automatic Video Aggregation in Social Media Networks	274
<i>Zhuhua Liao, Jing Yang, Chuan Fu, and Guoqing Zhang</i>	
Pedestrian Tracking Based on <i>Hidden-Latent</i> Temporal Markov Chain	285
<i>Peng Zhang, Sabu Emmanuel, and Mohan Kankanhalli</i>	
Motion Analysis via Feature Point Tracking Technology	296
<i>Yu-Shin Lin, Shih-Ming Chang, Joseph C. Tsai, Timothy K. Shih, and Hui-Huang Hsu</i>	
Traffic Monitoring and Event Analysis at Intersection Based on Integrated Multi-video and Petri Net Process	304
<i>Chang-Lung Tsai and Shih-Chao Tai</i>	
Baseball Event Semantic Exploring System Using HMM	315
<i>Wei-Chin Tsai, Hua-Tsung Chen, Hui-Zhen Gu, Suh-Yin Lee, and Jen-Yu Yu</i>	

Robust Face Recognition under Different Facial Expressions, Illumination Variations and Partial Occlusions	326
<i>Shih-Ming Huang and Jar-Ferr Yang</i>	
Localization and Recognition of the Scoreboard in Sports Video Based on SIFT Point Matching	337
<i>Jinlin Guo, Cathal Gurrin, Songyang Lao, Colum Foley, and Alan F. Smeaton</i>	
3D Model Search Using Stochastic Attributed Relational Tree Matching	348
<i>Naoto Nakamura, Shigeru Takano, and Yoshihiro Okada</i>	
A Novel Horror Scene Detection Scheme on Revised Multiple Instance Learning Model	359
<i>Bin Wu, Xinghao Jiang, Tanfeng Sun, Shanfeng Zhang, Xiqing Chu, Chuxiong Shen, and Jingwen Fan</i>	
Randomly Projected KD-Trees with Distance Metric Learning for Image Retrieval	371
<i>Pengcheng Wu, Steven C.H. Hoi, Duc Dung Nguyen, and Ying He</i>	
A SAQD-Domain Source Model Unified Rate Control Algorithm for H.264 Video Coding	383
<i>Mingjing Ai and Lili Zhao</i>	
A Bi-objective Optimization Model for Interactive Face Retrieval	393
<i>Yuchun Fang, Qiyun Cai, Jie Luo, Wang Dai, and Chengsheng Lou</i>	
Multi-symbology and Multiple 1D/2D Barcodes Extraction Framework	401
<i>Daw-Tung Lin and Chin-Lin Lin</i>	
Wikipedia Based News Video Topic Modeling for Information Extraction	411
<i>Sujoy Roy, Mun-Thye Mak, and Kong Wah Wan</i>	
Advertisement Image Recognition for a Location-Based Reminder System	421
<i>Siying Liu, Yiqun Li, Aiyuan Guo, and Joo Hwee Lim</i>	
Flow of Qi: System of Real-Time Multimedia Interactive Application of Calligraphy Controlled by Breathing	432
<i>Kuang-I Chang, Mu-Yu Tsai, Yu-Jen Su, Jyun-Long Chen, and Shu-Min Wu</i>	
Measuring Bitrate and Quality Trade-Off in a Fast Region-of-Interest Based Video Coding	442
<i>Salahuddin Azad, Wei Song, and Dian Tjondronegoro</i>	

Image Annotation with Concept Level Feature Using PLSA+CCA	454
<i>Yu Zheng, Tetsuya Takiguchi, and Yasuo Ariki</i>	
Multi-actor Emotion Recognition in Movies Using a Bimodal Approach	465
<i>Ruchir Srivastava, Sujoy Roy, Shuicheng Yan, and Terence Sim</i>	
Demo Session Papers	
RoboGene: An Image Retrieval System with Multi-Level Log-Based Relevance Feedback Scheme	476
<i>Huanchen Zhang, Haojie Li, Shichao Dong, and Weifeng Sun</i>	
Query Difficulty Guided Image Retrieval System	479
<i>Yangxi Li, Yong Luo, Dacheng Tao, and Chao Xu</i>	
HeartPlayer: A Smart Music Player Involving Emotion Recognition, Expression and Recommendation	483
<i>Songchun Fan, Cheng Tan, Xin Fan, Han Su, and Jinyu Zhang</i>	
Immersive Video Conferencing Architecture Using Game Engine Technology	486
<i>Chris Poppe, Charles-Frederik Hollemeersch, Sarah De Bruyne, Peter Lambert, and Rik Van de Walle</i>	
Author Index	489

A Generalized Coding Artifacts and Noise Removal Algorithm for Digitally Compressed Video Signals

Ling Shao¹, Hui Zhang², and Yan Liu³

¹ Department of Electronic & Electrical Engineering, The University of Sheffield, UK

² Department of Computer Science and Technology, United International College, China

³ Department of Computing, Hong Kong Polytechnic University, Hong Kong

Abstract. A generalized coding artifact reduction algorithm is proposed for a variety of artifacts, including blocking artifact, ringing artifact and mosquito noise. This algorithm does not require grid position detection and any coding parameters. All the filtering is based on local content analysis. Basically, the algorithm attempts to apply mild low-pass filtering on more informative regions to preserve the sharpness of the object details, and to apply strong low-pass filtering on less informative regions to remove severe artifacts. The size and parameters of the low-pass filters are changed continuously based on the entropy of a local region.

Keywords: Coding artifacts removal; noise reduction; entropy analysis; content adaptive filter.

1 Introduction

Most existing coding artifact reduction methods apply low-pass filters on block boundaries to remove blockiness [1-4]. Those algorithms all make the assumption that the block grid positions are available and are on the default 8x8 blocks. However, grid positions are not always available, especially when videos are rescaled after decoding. There are some grid position detection techniques, e.g. [5], available, but they are usually unreliable for scaled materials. It is desirable to design a coding artifact reduction algorithm without the knowledge of grid position or any other coding parameters. It can not only save the overload processing time of grid position detection, but also be a more generic approach for different kinds of materials.

Conventionally, different coding artifacts are removed separately. For example, in [6] blocking artifacts and ringing artifacts are detected and suppressed in a sequential way. Therefore, the reduction of one artifact may make another artifact more noticeable. We attempt to design an approach that targets different coding artifacts in a unified framework. In this scheme, we deem all the coding artifacts as noise, and try to differentiate noise from content. Noise in detailed regions is not as severe and noticeable as in flat regions. This is because of two reasons. The first reason is that an encoder usually allocates more bits for detailed regions than for flat regions, which results in less digital noise in detailed regions. The second reason is due to the masking effect of the human eyes, which makes noise less noticeable in detailed regions.

We employ entropy to indicate the information content of a local region. Local entropy has been used for artifacts reduction in [7, 8]. The drawback of [7] is that block grid detection is needed, and the shortcoming of [8] is that the footprint size of filters cannot be changed due to the principle of Least Squares trained filters.

In Section 2, we introduce the calculation of local entropy for evaluating the complexity and information content of an image region. In Section 3, we describe the algorithm for compression artifacts reduction based on local entropy analysis. Then, experimental results are discussed and evaluated in Section 4. Finally, the paper is concluded in Section 5.

2 Local Entropy Calculation

The performance of adaptive spatial filtering methods for image enhancement or digital artifacts reduction is often limited due to a lack of accuracy and robustness in the content classification scheme employed. Those methods usually tune the settings of the algorithms according to some local properties such as the mean gradient or variance of pixel intensities, which are heuristic and only effective for certain occasions. A general-purpose approach for region content classification is described in this section. Our approach is based on information theory, and specifically exploits the fact that a detailed region contains more information than a flat region. The information content of one region is quantized by the local entropy of the PDF of the pixel intensity distribution inside the region. The probability density function is approximated by the histogram of pixel intensities. We can observe that a detailed region has a flat and spread-out histogram, while the histogram of a smooth region only has a few peaks. Figure 1 shows the histograms of two regions, one on the eye and the other

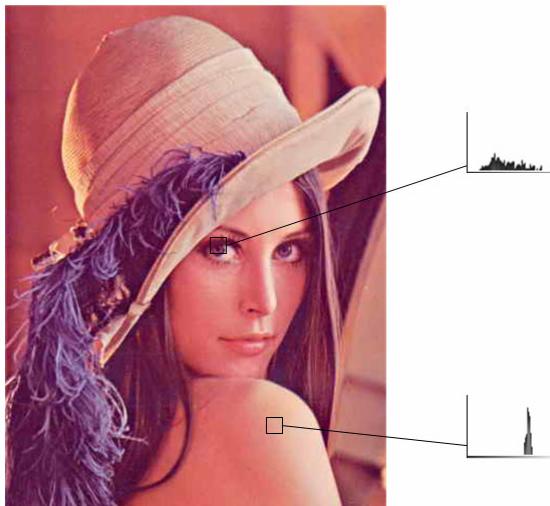


Fig. 1. Histograms of luminance intensity

on the shoulder of Lena (Lena is a JPEG decoded image). The histogram of the eye region is a lot more spread-out and distributed than the shoulder region. Note that the distribution of the histogram is dominated by the local structure of the region, i.e. noise and coding artifacts will not affect the overall distribution of the histogram.



Fig. 2. (a) An image with coding artifacts; (b) illustration of entropy value H : the brighter the pixel the higher the entropy value

The local entropy of a region can be defined as:

$$H_{D,R} = -\sum_i P_{D,R}(d_i) \log_2 P_{D,R}(d_i) \quad (1)$$

where $P_{D,R}(d_i)$ is the probability of descriptor D taking the value d_i in a local region R. The descriptors can be color, orientation, phase, etc. A scale invariant version of the local entropy has been used for salient region detection and object retrieval

in [9-11]. For simplicity and the consideration of the context of video processing, we employ luminance intensity as the descriptor. Therefore, the entropy calculation can be revised as follows:

$$H = -\sum_{i=0}^{255} P_R(i) \log_2 P_R(i) \quad (2)$$

where i indicates the intensity, R is a local region inside which the entropy is calculated, and P_R is the PDF of luminance intensity. According to the information theory, H has a higher value for a distributed histogram than a peaked one, i.e. the entropy value of a detailed region tends to be larger than a smooth region. As mentioned above, the entropy value is dependent on the information content or structure of the underlying region, and noise or coding artifacts would only deviate the entropy value in a small range.

Generally, the above entropy H can be used as an indicator of what kind of filtering should be applied on a region either for image enhancement or coding artifacts reduction. In the following artifacts reduction algorithm, the entropy H of each pixel is calculated on an $N \times N$ neighbourhood of that pixel. The window size for calculating entropy values should be carefully considered. If it is too small, it will not be representative for the local structure/content; otherwise it will be computationally too expensive. In our experiments, a window size of 7×7 is used, since it is approximately the size of a decent low-pass filter for compression artifacts removal. Figure 2 illustrates the entropy value H of each pixel in a decompressed image. The brighter the pixel in Figure 2(b), the higher entropy value the corresponding pixel in Figure 2(a) has. It can be seen that entropy H nicely represents the smooth transition from detailed regions to flat regions, and it is not affected by digital artifacts. The entropy map in Figure 2(b) is different to an edge map, because not only edges have high values but also textures and fine details. The entropy value is also more robust to coding artifacts, since blocking artifacts would be detected as edges.

3 Compression Artifacts and Noise Removal

The aim of a good coding artifacts reduction algorithm is to preserve the sharpness of object details and to remove the coding artifacts. Therefore, the kernel size and parameters of the low-pass filters should be dependent on local structure. The more detail a region has, the less smoothing should be applied; and the less detail a region contains, the more smoothing can be used.

The entropy H of each pixel location is first cropped to be in the range of $[H_{\min}, H_{\max}]$. The kernel size S of the low-pass filter used is dependent on the value of H . S should be enlarged with the decrease of H . Both linear and non-linear relationships can be used between S and H . Non-linear relationships can theoretically lead to more optimized performance. However, there is no theory to automatically define the non-linear relationship and a non-linear relationship is less efficient computationally. For simplicity, linear relationship as illustrated in Figure 3 is adopted. It can also be expressed in the following equation:

$$S = S_{\min} + (S_{\max} - S_{\min}) \frac{H_{\max} - H}{H_{\max} - H_{\min}} \quad (3)$$

where $[S_{\min}, S_{\max}]$ is the range of S . Therefore, the kernel size S is small when entropy H is large, and is large when H is small.

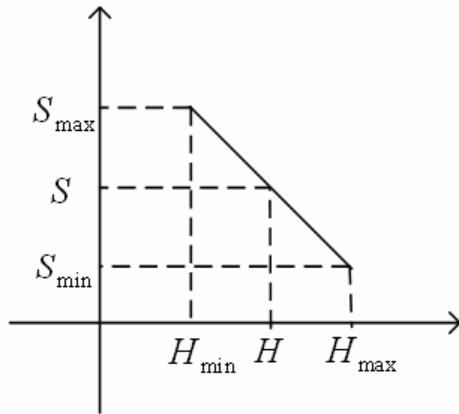


Fig. 3. The relationship between S and H

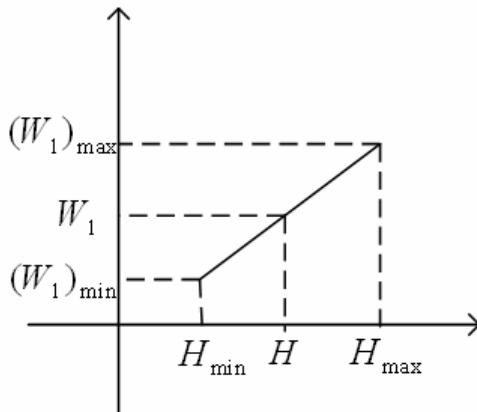


Fig. 4. The relationship between W_1 and H

The low-pass filter we use is a bilateral-like filter, i.e. only pixels whose intensity values are within a certain range of the intensity of the current pixel are included for filtering. The threshold of the difference between intensities of the current pixel and pixels in the kernel is also dependent on the entropy value H . A linear relationship between the threshold and the entropy value is also used here. The boundaries of the linear curve are chosen based on manual tuning. A smaller threshold is used for detailed regions and a larger threshold is used for flat regions, which enables the algorithm to remove very strong blocking artifacts in flat regions.

The parameters of the low-pass filters should also be changed based on entropy H . Here we use a simple version of the filter kernel, i.e. there are only two parameter values: one for the central pixel and the other for the remaining neighbouring pixels.

W_1 represents the filter coefficient of the central pixel, and W_2 is the filter coefficient for the remaining pixels. If the number of pixels in the kernel after thresholding is $n+1$, the relationship between W_1 and W_2 should be as follows:

$$W_1 + nW_2 = 1 \quad (4)$$



Foreman



Girlsea



Fashion

Fig. 5. Snapshots of test sequences

If we change W_1 , W_2 will be changed accordingly. We want to employ a stronger low-pass filter in flat regions than in detailed regions. We define the relationship between W_1 and H as follows:

$$W_1 = (W_1)_{\min} + [(W_1)_{\max} - (W_1)_{\min}] \frac{H - H_{\min}}{H_{\max} - H_{\min}} \quad (5)$$

where $[(W_1)_{\min}, (W_1)_{\max}]$ is the range of W_1 . Figure 4 also illustrates the linear relationship between W_1 and H .

The continuously changing of the size and parameters of the filter kernel makes the transition of processing from detailed regions to flat regions smooth. It overcomes the abrupt change problem in algorithms using discrete modes.

In our implementation, simpler solutions, such as linear transforms, are adopted to meet our requirements of real-time processing in the scenario of consumer electronics products, e.g. TV.

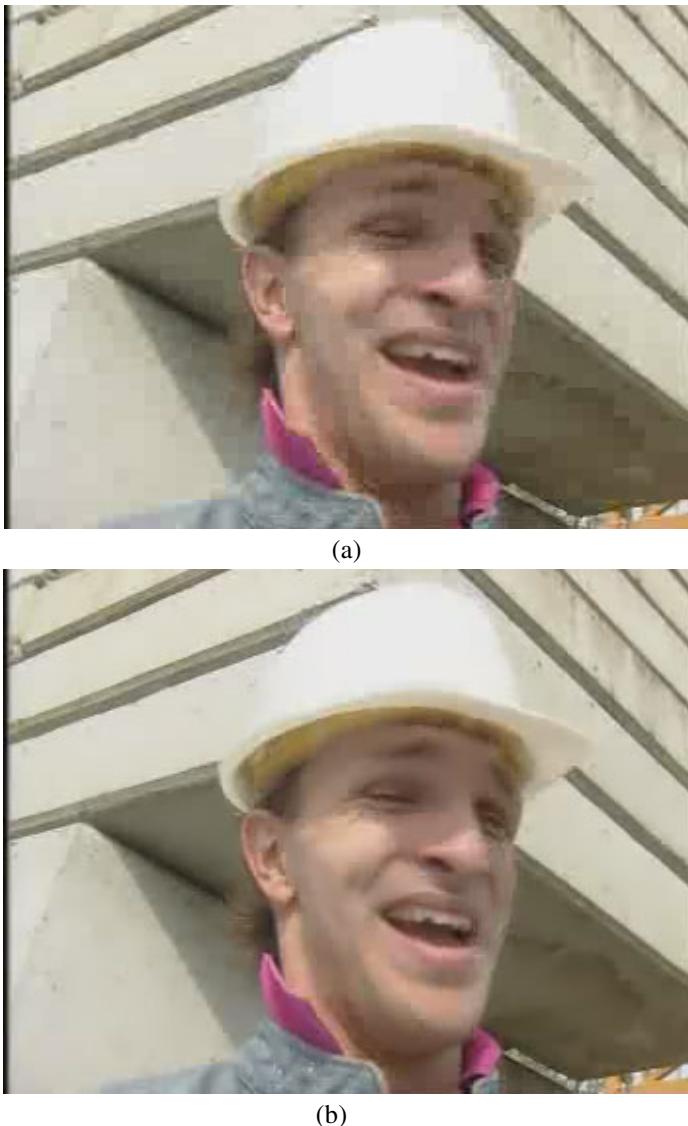
4 Results and Evaluation

The proposed algorithm is applied on MPEG-4 compressed materials with different compression rates. The snapshots of the test sequences are shown in Figure 5. Figure 6 shows the output after coding artifacts reduction of the Foreman sequence. It is easy to see that the severe blocking artifacts on flat regions have been nicely removed, and the ringing artifacts around object edges are reduced without blurring object details.

For objective evaluation, we calculate the Peaked Signal to Noise Ratio (PSNR) between the original uncompressed sequences and the output sequences after applying the proposed artifacts reduction algorithm on the decompressed sequences containing coding artifacts. The PSNR scores for different sequences with different compression rates are given in Table 1. For benchmarking, the results of two state-of-the-art artifacts reduction methods proposed in [12] and [7] are also shown. Both of the methods require block grid detection, and for them to work properly the test sequences are not scaled intentionally in the experiments. The comparison shows that our proposed algorithm outperforms the other two methods in PSNR, though no block grid detection is needed in our algorithm.

Table 1. PSNR Comparison the Test Sequences

Sequence	Bit-rate (Mbit/s)	Proposed	Ref [12]	Ref [7]
Foreman (CIF)	0.1	29.98	29.75	29.88
	0.2	32.07	31.38	31.98
	0.5	34.68	32.87	34.59
Girlsea (SD)	1.0	33.41	31.71	32.80
	2.0	36.05	33.28	35.43
	3.0	37.70	33.99	37.09
Fashion (HD)	4.0	40.77	40.18	40.25
	6.0	41.72	41.05	41.17
	8.0	42.16	41.47	41.68



(b)

Fig. 6. Results of the Foreman sequence: (a) before processing; (b) after processing using the proposed method

5 Conclusion

A generalized approach is proposed for reducing various coding artifacts. The size and coefficients of the low-pass filters are controlled by local entropy continuously. The method can be reliably applied on scaled materials, because no block grid detection is required. Experimental results show that the proposed algorithm performs better than other methods that require block grid detection.

References

- [1] Chen, T., Wu, H.R., Qiu, B.: Adaptive postfiltering of transform coefficients for the reduction of blocking artifacts. *IEEE Transactions on Circuits and Systems for Video Technology* 11(5), 584–602 (2001)
- [2] Kim, S., Yi, J., Kim, H., Ra, J.: A deblocking filter with two separate modes in block-based video coding. *IEEE Transactions on Circuits and Systems for Video Technology* 9 (1999)
- [3] Pan, F., Lin, X., Rahardja, S., Lin, W., Ong, E., Yao, S., Lu, Z., Yang, X.: A locally-adaptive algorithm for measuring blocking artifacts in images and videos. In: *IEEE International Symposium on Circuits and Systems* (2004)
- [4] Tai, S., Chen, Y., Sheu, S.: Deblocking filter for low bit rate MPEG-4 video. *IEEE Transactions on Circuits and Systems for Video Technology* 15 (2005)
- [5] Kireenko, I., Muijs, R., Shao, L.: Coding artifact reduction using non-reference block grid visibility measure. In: *Proc. IEEE International Conference on Multimedia and Expo*, Toronto, Canada (July 2006)
- [6] Kong, H.-S., Vetro, A., Sun, H.: Edge map guided adaptive post-filter for blocking and ringing artifacts removal. In: *Proc. IEEE International Symposium on Circuits and Systems*, vol. 3, pp. 929–932 (May 2004)
- [7] Shao, L., Kireenko, I.: Coding artifact reduction based on local entropy analysis. *IEEE Trans. Consumer Electronics* 53(2), 691–696 (2007)
- [8] Shao, L.: Unified compression artifacts removal based on adaptive learning on activity measure. *Digital Signal Processing* 17(6), 1065–1070 (2007)
- [9] Shao, L., Kadir, T., Brady, M.: Geometric and Photometric Invariant Distinctive Regions Detection. *Information Sciences* 177(4), 1088–1122 (2007)
- [10] Shao, L., Brady, M.: Specific object retrieval based on salient regions. *Pattern Recognition* 39(10), 1932–1948 (2006)
- [11] Shao, L., Brady, M.: Invariant salient regions based image retrieval under viewpoint and illumination variations. *Journal of Visual Communication and Image Representation* 17(6), 1256–1272 (2006)
- [12] Kireenko, I.: Reduction of coding artifacts using chrominance and luminance spatial analysis. In: *Proc. IEEE International Conference on Consumer Electronics*, Las Vegas, USA (January 2006)

Efficient Mode Selection with BMA Based Pre-processing Algorithms for H.264/AVC Fast Intra Mode Decision

Chen-Hsien Miao and Chih-Peng Fan^{*}

Department of Electrical Engineering,
National Chung Hsing University,
Tai-chung 402, Taiwan, R.O.C.

g9764429@mail.nchu.edu.tw, cpfan@dragon.nchu.edu.tw

Abstract. In a H.264/AVC intra-frame encoder, the complicated computations for the mode decision cause the difficulty in real-time applications. In this paper, we propose an efficient fast algorithm, which is called Block Matching Algorithm (BMA), to predict the best direction mode for the fast intra mode decision. The edge detective technique can predict luma-4x4, luma-16x16, and chroma-8x8 modes directly. The BMA method uses the relations between the current block and the predictive block to predict edge directions. We can partition the intra prediction procedure into two steps. At the first step, we use the pre-processing mode selection algorithms to find the primary mode which can be selected for fast prediction. At the second step, the selected fewer high-potential candidate modes are applied to calculate the RD cost for the mode decision. The encoding time is largely reduced, and meanwhile we also maintain the same video quality. Simulation results show that the proposed BMA method reduces the encoding time by 75%, and requires bit-rate increase about 2.5% and peak signal-to-noise ratio (PSNR) decrease about 0.07 dB in QCIF and CIF sequences, compared with the H.264/AVC JM 14.2 software. Our methods can achieve less PSNR degradation and bit-rate increase than the previous methods with more encoding time reduction.

Keywords: H.264/AVC, Fast algorithm, Intra prediction, Intra-mode decision, Rate-distortion optimization (RDO).

1 Introduction

To develop the H.264/Advanced Video Coding (AVC), the ITU-T Video Coding Experts Group (VCEG) and the Moving Picture Experts Group (MPEG) formed the Joint Video Team (JVT) in 2001 to develop the new video coding standard [1]. The H.264/AVC video compression can reduce bit-rate by about 60% over MPEG-2, and bit-rate saving by about 40% over MPEG-4 at the same video quality. In other words, the H.264 has better compression ratio than the previous video coding standards. Although the H.264/AVC scheme is outstanding, the computational complexity of H.264 is considerably complex. When the H.264 compression is applied in order to

^{*} Corresponding author.

reduce the encoding complexity but almost maintain the suitable coding efficiency, achieving real-time applications is very important. By studying the H.264 system, the inter-prediction and intra-prediction require a large number of computations, such as the variable block size motion estimation, the intra-mode prediction, and the intra/inter-mode decision. According to the above mentions, the acceleration methods for intra and inter predictions are key technologies to the success of the H.264 real-time applications. To achieve the best coding performance, the rate-distortion optimization (RDO) for the mode decision is an important step. The H.264 reference software (JM) recommends two mode decision schemes, i.e. low and high complexity mode decisions, and the rate-distortion (RD) cost function is used for both mode decisions. The RD cost function for both mode decisions can discover the best mode which has the minimum RD cost value. The high complexity mode decision requires a large number of computations but it leads to the best performance, so it is not suitable for real-time applications. To conquer the computational complexity, an efficient and fast intra-mode selection algorithm must be developed to reduce the calculations and speed up the encoding time in software for the H.264/AVC real-time realization.

Recently, many fast intra-mode decision algorithms have been proposed to reduce mode decision computations and these fast methods have reduced the encoding time effectively. Huang *et al.* [4] used the sub-sampled resolution for the RD cost function and applied the context-based most probable modes table to skip unlikely candidate modes, and then the encoding time was reduced by about 50%. Pan *et al.* [5] used the Sobel operator to detect edge angles of 4x4 blocks and 16x16 macroblocks, and this fast algorithm reduced candidate modes to save computations and the encoding time. Cheng *et al.* [6] proposed the three-step fast algorithm to simplify the cost generation and the mode decision. To reduce the mode computation for the mode decision, the fast method in [6] only computed the RD cost value of the neighboring modes which were close to the primary mode, where the primary mode had the small RD cost value, so it could skip some modes, whose RD cost values were larger than the primary mode. Wang *et al.* [7] and Tsai *et al.* [8] supported regular and independent pre-processing algorithms to detect edge directions inside the current block directly, and they could find the dominant direction by their edge detection techniques. Tsai *et al.* [8] proposed two fast algorithms, i.e. the pixel-based direction detection (PDD) and the subblock-based direction detection (SDD).

In this paper, we propose an effective pre-processing fast algorithm, which is called Block Matching Algorithm (BMA). To find the primary direction, the fast algorithm, BMA, uses the relation of the current and predictive blocks after intra prediction. We use the mode selection algorithm to approach mode reduction in the mode decision and speed up the encoding time. Simulation results show that the proposed algorithm almost outperforms the previous methods. Compared with the H.264/AVC JM 14.2 software, the proposed BMA method reduces the encoding time by 75%, requires the PSNR degradation about 0.07dB in QCIF and CIF formats, and pays the bit-rate increase about 2.5% in QCIF and CIF formats.

The rest of the paper is organized as follows. In Section 2, the H.264/AVC intra prediction and the mode decision are introduced. The proposed BMA based algorithm is described in Section 3. Experimental results, comparisons and the mode reduction in RDO calculations among different algorithms are presented in Section 4. Finally, a conclusion is stated at the end of the paper.

2 Overview of Intra Prediction and Mode Decision

Intra prediction plays an important role in H.264/AVC intra frame encoding procedure, and it processes three kinds of blocks, i.e. luminance intra-4x4, luminance intra-16x16, and chrominance intra-8x8 blocks. There are nine intra prediction modes in luminance 4x4 blocks, four modes in the luminance 16x16 or chrominance 8x8 blocks. In the intra-frame encoder of H.264/AVC, the RD cost requires most computations and spends much time in the mode decision, and then all nine intra modes must be calculated with the cost function, and the minimum cost for the best mode can be found. Thus, to reduce intra mode computation in the cost function, we can use some fast intra-mode decision methods before the mode decision. The RDcost equation is shown as follows.

$$RDcost = Distortion + \lambda \times Rate, \quad (1)$$

where “Distortion” denotes the sum of the absolute transform difference (SATD) or the sum of the square difference (SSD) between the current block and the predicted block, which has been computed by the intra prediction, “Rate” means the bit-rate, which is estimated by the number of bits required to encode the mode information and the residue, and “ λ ” is the Lagrangian coefficient, which is dependent on the quantization parameter (or QP).

3 Proposed Intra-mode Pre-processing Selection Algorithm

In this section, we propose an efficient pre-processing fast method to approach the modes reduction and decrease the computational complexity in the cost generation and the mode decision. In the original H.264 JM intra frame encoding procedure, the mode decision will compute RD cost of all combinational intra modes for luma and chroma components, and the computation becomes very large, and then it slows down the encoding time, so we can insert additional pre-processing technique before the mode decision to skip low potential intra modes. By the technique, we can predict the edge direction of the luma and chroma current blocks independently, and then we find the primary direction as the dominant mode. If other modes are so different to the dominant mode, we don’t consider them and don’t compute their RD costs, so the computational complexity will be reduced. The details of the proposed fast algorithm will be discussed as follows.

We propose the fast mode selection algorithm, which is called Block Matching Algorithm (BMA) with a two-stage intra prediction, where Fig. 1 shows the proposed coding scheme. We can use the relationship between the current block and predictive block to find the primary mode directly. We can also use two methods to do the BMA, where one is based on the pixel method, called Pixel-Based Block Matching Algorithm (PBMA), and the other is based on the block method, called Block-Based Block Matching Algorithm (BBMA). In Fig. 1, for the luma-4x4 prediction procedure, the BMA uses the predictive block from the intra prediction at the first stage and the current block from the current frame to predict the primary mode, and then it sends the primary mode information to the intra prediction at the second stage, and the intra prediction at the second stage will compute two predictive blocks of two neighboring modes. Finally, we send the primary and DC modes generated from BMA and two neighboring modes generated from the second-stage intra prediction to

the cost generation for the mode decision. By the proposed fast algorithm, the intra prediction modes will be reduced in the mode decision.

In Wang [7] and Tsai [8], the major four directions, i.e. vertical, horizontal, diagonal down-left, and diagonal down-right directions, are very distinctive, so their edges can be detected easily. These directions will be searched firstly, and then we can find the primary direction by a simple way before the mode decision, and only few modes are sent to compute the RD cost if the primary mode is decided. The details of the proposed fast mode selection algorithms are shown as follows:

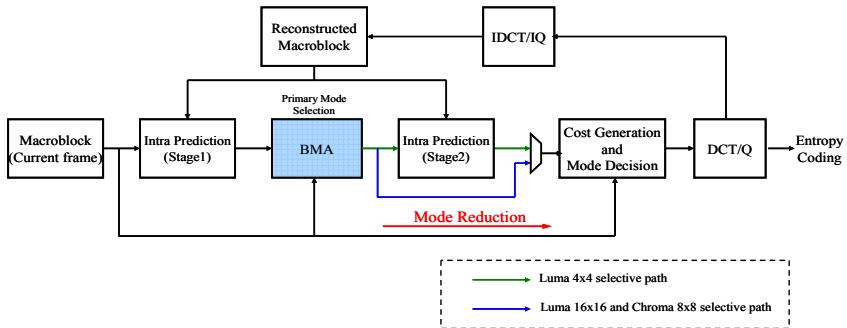


Fig. 1. The coding scheme of the H.264/AVC intra encoder with the proposed BMA method

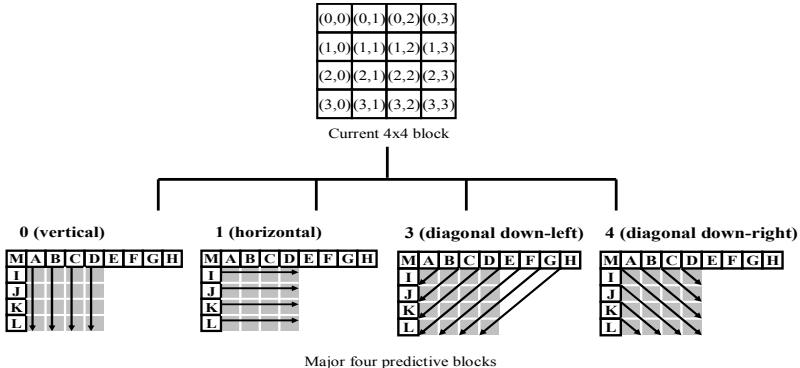
3.1 Pixel-Based Block Matching Algorithm (PBMA)

For the PBMA method, we calculate the pixel-domain sum-absolute-difference (PSAD) between the current block and predictive blocks. The pixel-domain SAD (PSAD) calculation of one 4x4 block is shown as follows:

$$PSAD = \sum_{i=0}^3 \sum_{j=0}^3 |cur(x+i, y+j) - pred(x+i, y+j)|, \quad (2)$$

where $cur(x,y)$ means the pixel value at the (x,y) position in the current block, and $pred(x,y)$ means the pixel value at the (x,y) position in the predictive block. For the mode detection of luminance 4x4 blocks, we divide the fast intra prediction into two stages firstly, and then the PBMA scheme is located between the two-stage intra predictions.

In the first-stage intra prediction, four directional modes, which are vertical (mode 0), horizontal (mode 1), diagonal down-left (mode 3), diagonal down-right (mode 4) and non-directional DC modes (mode 2), are predicted, and then five kinds of predictive intra blocks are obtained, and the major four directional predictive blocks are sent to the PBMA scheme to compute the PSAD with the current block, where the PBMA schemes between the current block and four directional predictive blocks are shown in Fig. 2. After computations by block matching, we can find the best matched predictive block which has the smallest PSAD value, and then the primary mode will be decided.

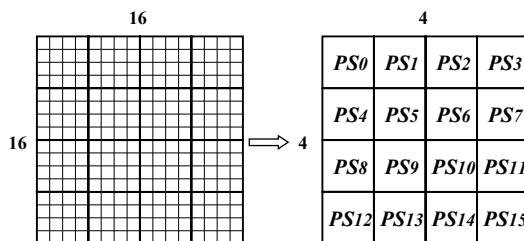
**Fig. 2.** Compute PSAD for a 4x4 luma block

In the second-stage intra prediction, two neighboring modes of the primary directional mode are also calculated. At the second stage, we will choose three candidate modes, i.e. one is the primary mode, and the other two are neighboring modes. Then three directional modes and one non-directional DC mode will be computed with RDcost values and the best mode will be decided. Therefore, there are four modes instead of nine in the best case. But In the worst case, if the smallest PSAD can not decide the primary mode, the mode decision will compute the RD cost of five higher potential modes which are mode 0 to mode 4.

For a 16x16 luminance block, it can be divided into sixteen 4x4 subblocks initially, and then the average of each 4x4 subblock is computed to get an average pixel, and a 16x16 block can be replaced with a 4x4 average block, which is shown in Fig. 3. For the average of each 4x4 subblock, all sixteen average pixels, PS_k , of a 16x16 luma block are expressed as follows.

$$PS_k = \frac{\sum_{i=0}^3 \sum_{j=0}^3 f_k(i, j)}{16}, \quad (3)$$

for $k = 0, 1, 2, \dots, \text{and } 15$, where $f_k(i, j)$ means the pixel value at the (i, j) position in the k -th luma-4x4 subblock. Then we can use (3) to calculate the average-block PSAD between the 4x4 average current block and the 4x4 average predictive block.

**Fig. 3.** One 16x16 luma block can be replaced with one 4x4 average block

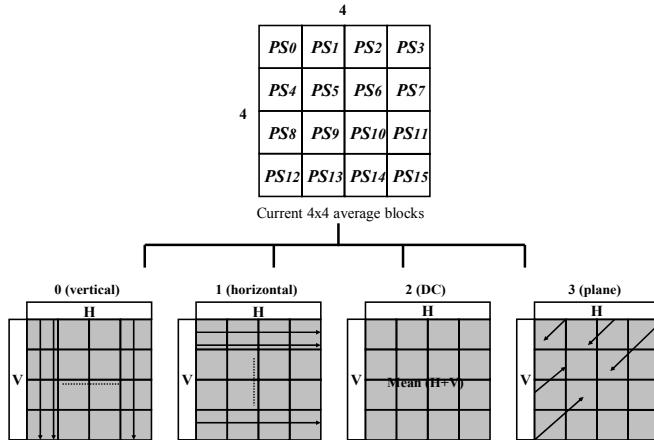


Fig. 4. Compute the average-block PSAD for a 16x16 luma block

In the luma-16x16 mode selection procedure, we only have one stage intra prediction. The intra prediction unit computes total four predictive intra blocks, which are vertical (mode 0), horizontal (mode 1), DC (mode 2) and plane modes (mode 3), and we use the proposed method to select all four modes. The predictive schemes between the luma-16x16 current block and total four predictive blocks are shown in Fig. 4. After comparisons, the predictive mode with the smallest SAD is selected as the primary mode.

For one 8x8 chrominance block, it can be divided into sixteen 2x2 subblocks, and then the average of each 2x2 subblock is computed, and one 8x8 block can be replaced with one 4x4 average block. For the average of each 2x2 subblock, all sixteen average pixels, PE_k , of an 8x8 chroma block are expressed as follows.

$$PE_k = \frac{\sum_{i=0}^1 \sum_{j=0}^1 f_k(i, j)}{4}, \quad (4)$$

for $k = 0, 1, 2, \dots, 15$, where $f_k(i, j)$ means the pixel value at the (i, j) position in the k -th chroma 2x2 subblock. We can use the 4x4 average block to do the PBMA method that is similar to above being mentioned.

3.2 Block-Based Block Matching Algorithm (BBMA)

In this part we propose another kind of BMA method, which is called Block-based Block Matching Algorithm (BBMA). As to the BBMA method, firstly we divide a luma 4x4 block into nine 2x2 subblocks, which are indicated with S_0, S_1, \dots, S_7 as shown in Fig. 5. S_0, S_2, S_5 and S_7 are corner subblocks in a luma 4x4 block, and S_1, S_3, S_4 and S_6 are inner subblocks which overlap S_0, S_2, S_5 , and S_7 subblocks. Each subblock value, S_k , can be shown as

$$S_k = f_k(0,0) + f_k(0,1) + f_k(1,0) + f_k(1,1), \quad (5)$$

for $k = 0, 1, 2, \dots$, and 8, where $f_k(i, j)$ means the pixel value at the (i, j) position in the k -th luma 2×2 subblock. Finally, we calculate the block-based SAD (BSAD) between the current and the predictive blocks. The BSAD of one 4×4 luma block can be shown as

$$BSAD = \sum_{k=0}^7 |S_k(\text{current}) - S_k(\text{predictive})|. \quad (6)$$

The luma- 4×4 mode selection procedure is similar to the PBMA method. Therefore, there are only four modes instead of nine, which will be chosen for the mode decision. But in the worst case, if the primary mode can not be decoded by the smallest BSAD, the mode decision will compute the RD costs of the mode 0 to mode 4.

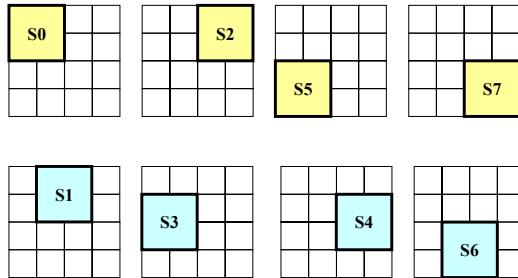


Fig. 5. Up: the four corner subblocks; Down : the four inner subblocks

For the luma intra- 16×16 and chroma- 8×8 mode selections, we can divide one macroblock into eight 8×8 subblocks and divide one 8×8 block into eight 4×4 subblocks in the luma- 16×16 and chroma- 8×8 processes, respectively. Then we accumulate all pixels inside subblocks. Each subblock value SS_k in a luma 16×16 block and the subblock value SE_k in a chroma 8×8 block can be described as

$$SS_k = \sum_{i=0}^7 \sum_{j=0}^7 f_k(i, j), \quad (7)$$

$$SE_k = \sum_{i=0}^3 \sum_{j=0}^3 f_k(i, j), \quad (8)$$

for $k = 0, 1, 2, \dots$, and 7, where $f_k(i, j)$ means the pixel value at the (i, j) position in the k -th luma 8×8 or chroma 4×4 subblock as shown in Fig. 5. We compute the BSADs of eight subblocks between the current block and four predictive blocks, the procedure is similar to above being mentioned, and then we find the primary mode which has the smallest SAD.

For the chroma- 8×8 mode selection procedure with the PBMA or BBMA method, we divide the process into two steps. At the first step, we use one chroma component of U or V to compute three directional modes, which are vertical (mode 1), horizontal (mode 2) and plane (mode 3) modes, and then we find the primary directional mode which has the smallest SAD. At the second step, we use both chroma components of U and V to compute the primary directional mode and the non-directional DC mode, and find the final primary mode, and the procedure.

For both luma-16x16 and chroma-8x8 mode selection processes, when the primary mode is decided, we can finally send the primary mode to the cost generation and mode decision unit, which means that the primary mode is the best mode. For example, when the vertical mode has the smallest SAD, it is selected and sent to the mode decision. The processes of the other three modes are similar to those of the vertical mode. In summary, only one mode is computed with the RD cost computation instead of four, so we can obtain the best mode by using the algorithm directly.

4 Experimental Results, Comparisons and Discussions

In order to verify the video quality with the proposed fast algorithms, we evaluate the encoding time reduction, the bit-rate (BR) increase, and the peak-signal-to-noise ratio (PSNR) degradation in the H.264 intra encoder. The intra prediction scheme is simulated by the proposed BMA algorithms. The proposed fast mode decision algorithms are implemented on the JM 14.2 software platform provided by JVT. For a fair comparison, we also implement and simulate all the compared algorithms on the same JM 14.2 software and the same personal computer platform, where the CPU operates at the 2.26GHz working frequency. In our experimental environment, configurations of profile and level are baseline and 4.0, and all of the frames are encoded by using I-frame coding, and then each video sequence contains 300 frames, and the high complexity RDO is enabled. To compare the rate distortion and the quality performance of the proposed algorithm with the others, the PSNR and the bit-rate are measured by using Bjontegaard's method [9] for the quantization parameters (QPs) assigned at 28, 32, 36, and 40, respectively. Tables 1 and 2 show the BMA method in QCIF and CIF sequences respectively, where Δ PSNR denotes the average difference of the peak signal-to-noise ratio, Δ BR indicates the average bit-rate increase, and Δ TIME indicates the average time saving in coding process. The performance indices for comparisons are defined as

$$\Delta \text{PSNR} = \text{PSNR}_{\text{method}} - \text{PSNR}_{\text{ref}} , \quad (9)$$

$$\Delta \text{BR} = \frac{\text{BR}_{\text{method}} - \text{BR}_{\text{ref}}}{\text{BR}_{\text{ref}}} \times 100\% , \quad (10)$$

$$\Delta \text{TIME} = \frac{\text{TIME}_{\text{method}} - \text{TIME}_{\text{ref}}}{\text{TIME}_{\text{ref}}} \times 100\% . \quad (11)$$

4.1 Comparison of Proposed and Previous Algorithms

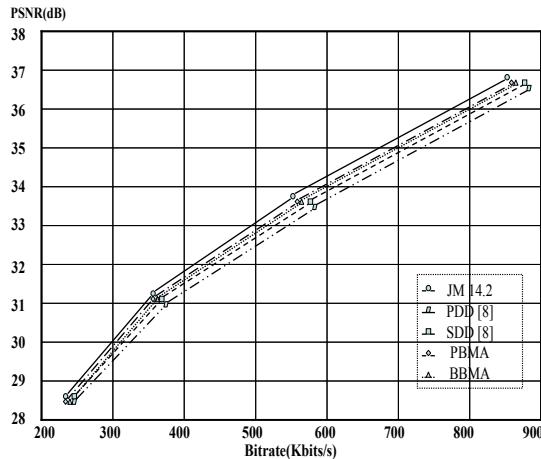
From Tables 1 and 2, the proposed PBMA method is also better than the previous methods [8] in encoding timing reduction and quality performance. The PBMA method decreases about 0.06dB in PSNR and increases about 1.7% in BR in the QCIF sequences, and decreases about 0.07dB in PSNR and increases about 2.4% in BR in the CIF sequences. The quality performance of BBMA method is better than the PDD [8] method and close to the SDD [8] method. The BBMA method decreases about 0.06dB in PSNR and increases about 2.4% in BR in the QCIF sequences, and decreases about 0.07dB in PSNR and increases about 3.2% in BR in the CIF sequences. The encoding time of all BMA methods is reduced by about 15% more

Table 1. Comparison results with QCIF sequences for the proposed BMA based methods

Sequence	PDD [8]			SDD [8]			PBMA			BBMA		
	Δ PSNR [dB]	Δ BR [%]	Δ TIME [%]	Δ PSNR [dB]	Δ BR [%]	Δ TIME [%]	Δ PSNR [dB]	Δ BR [%]	Δ TIME [%]	Δ PSNR [dB]	Δ BR [%]	Δ TIME [%]
Foreman	-0.10	4.1848	-60.5386	-0.06	2.8415	-59.3750	-0.05	2.0052	-75.9643	-0.05	2.9246	-75.8798
News	-0.08	3.8427	-61.5823	-0.10	2.4780	-59.1478	-0.06	2.3113	-75.6954	-0.06	2.9925	-75.4658
Container	-0.09	2.4784	-60.9491	-0.08	3.9080	-59.6619	-0.06	1.9008	-76.1440	-0.06	2.5548	-76.0004
Silent	-0.14	3.6964	-62.1303	-0.07	3.1746	-59.7416	-0.08	1.1585	-75.5986	-0.08	1.6783	-75.6535
Coastguard	-0.08	1.8090	-62.9801	-0.07	1.3634	-61.2660	-0.05	1.1948	-75.3485	-0.04	1.9094	-75.3556
Average	-0.10	3.2023	-61.6361	-0.08	2.7531	-59.8385	-0.06	1.7141	-75.7502	-0.06	2.4119	-75.6710

Table 2. Comparison results with CIF sequences for the proposed BMA based methods

Sequence	PDD [8]			SDD [8]			PBMA			BBMA		
	Δ PSNR [dB]	Δ BR [%]	Δ TIME [%]	Δ PSNR [dB]	Δ BR [%]	Δ TIME [%]	Δ PSNR [dB]	Δ BR [%]	Δ TIME [%]	Δ PSNR [dB]	Δ BR [%]	Δ TIME [%]
Paris	-0.09	2.6935	-63.7476	-0.07	1.8865	-60.7909	-0.05	1.6316	-77.8742	-0.04	2.4297	-77.7312
Mobile	-0.15	1.8152	-65.6203	-0.12	1.5322	-61.7547	-0.09	1.0536	-79.6008	-0.09	1.5628	-79.4445
Tempeste	-0.16	2.3864	-63.8874	-0.11	1.7495	-60.8219	-0.09	0.8890	-78.3176	-0.09	1.3863	-78.1749
Hall Monitor	-0.06	4.9202	-58.4979	-0.10	3.6453	-57.8690	-0.04	3.3568	-75.0374	-0.04	4.4224	-74.8282
Mother and Daughter	-0.12	6.9031	-56.5848	-0.12	7.1485	-55.3575	-0.06	5.2641	-73.4834	-0.07	6.1911	-73.1249
Average	-0.12	3.7437	-61.6676	-0.10	3.1924	-59.3188	-0.07	2.4390	-76.8627	-0.07	3.1985	-76.6607

**Fig. 6.** RD curve for Silent sequence in QCIF

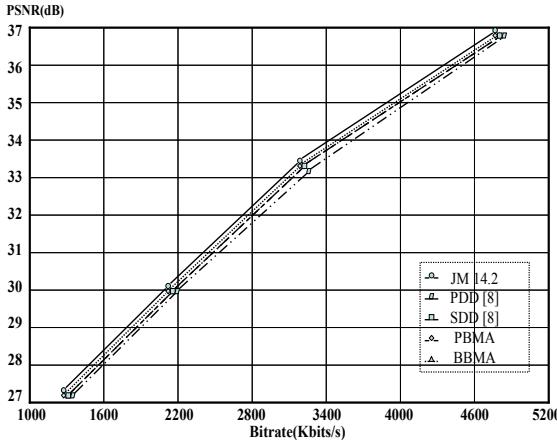


Fig. 7. RD curve for Tempete sequence in CIF

than [8] while they are maintaining better quality performance. Figs. 6 and 7 show the RD curves achieved by the proposed BMA method and the compared methods with QCIF (4:2:0) and CIF (4:2:0) sequences, respectively.

4.2 Discussion for Encoding Time Reduction

The reason which influences the encoding time is the mode reduction. We will reduce more encoding time if we choose fewer modes in the RDO calculation. Our proposed methods almost select four modes in the mode decision, so the mode selection decreases about 50% more than the original luma-4x4 procedure. The total combinational modes of both luma and chroma components for one macroblock are 592, and the computational complexity is very large when we use high complexity RDO calculation. After the proposed pre-processing methods are added to the H.264 system, we can efficiently reduce the combinational modes in the mode decision. The average selective modes in PDD [8] are 140.64, so it reduces combinational modes by about 76.24% in the RDO calculation. The average selective modes in SDD [8] are 178.08, so it reduces combinational modes by about 69.92% in the RDO calculation. The average selective modes in the PBMA and BBMA methods are 67.24 and 69.48, so they reduce combinational modes by about 88.64% and 88.26% in the RDO calculation, respectively. In summary, the encoding time reduction of the proposed algorithms is more than that of the previous work in [8].

5 Conclusion

A lot of computations for the mode decision cause the difficulty for real-time applications in H.264/AVC. In order to reduce predictive modes and speed up the encoding time in the mode decision, we can estimate the edge directions in the blocks by fast prediction schemes. The paper states an efficient and simple intra-mode decision algorithm for the H.264/AVC encoder, and the main purpose is to reduce some low potential modes in the mode decision and speed up the encoding time with

the same video quality. Because the proposed fast algorithm is independent and it doesn't use complicated information and technique to detect edge directions, we can directly convert soft IP to hard IP in the future. Simulation results show that the BMA method reduces the encoding time by 75%, and requires bit-rate increase about 2.5% and peak signal-to-noise ratio (PSNR) decrease about 0.07 dB in QCIF and CIF sequences, compared with the H.264/AVC JM 14.2 software. The proposed method greatly reduces the computational complexity of the cost function in the H.264 intra-mode decision but also maintains better video quality than the previous work in [8].

Acknowledgement

This work was supported by the National Science Council, Taiwan, R.O.C., under Grant NSC97-2221-E-005-088-MY2.

References

1. Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec.H.264 ISO/IEC 14496-10 AVC), Joint Video Team, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-G050 (2003)
2. Richardson, I.E.G.: H.264 and MPEG-4 Video Compression – Video Coding for Next - generation Multimedia. John Wiley & Sons, Chichester (2004)
3. Wiegand, T., Sullivan, G.J., Bjontegaard, G., Luthra, A.: Overview of the H.264/ AVC Video Coding Standard. IEEE Transactions on Circuits and Systems for Video Technology 13(7), 560–576 (2003)
4. Huang, Y.W., Hsieh, B.Y., Chen, T.C., Chen, L.G.: Analysis, Fast Algorithm, and VLSI Architecture Design for H.264/AVC Intra-frame Coder. IEEE Transactions on Circuits and Systems for Video Technology 15(3), 378–401 (2005)
5. Pan, F., Lin, X., Rahardja, S., Lim, K.P., Li, Z.G., Wu, D., Wu, S.: Fast Mode Decision Algorithm for Intra-prediction in H.264/AVC Video Coding. IEEE Transactions on Circuits and Systems for Video Technology 15(7), 813–822 (2005)
6. Cheng, C.C., Chang, T.S.: Fast Three Step Intra Prediction Algorithm for 4x4 blocks in H.264. IEEE ISCAS 2, 1509–1512 (2005)
7. Wang, J.C., Wang, J.F., Yang, J.F., Chen, J.T.: A Fast Mode Decision Algorithm and Its VLSI Design for H.264/AVC Intra Prediction. IEEE Transactions on Circuits and Systems for Video Technology 17(10), 1414–1422 (2007)
8. Tsai, A.C., Wang, J.F., Yang, J.F., Lin, W.G.: Effective Subblock-Based and Pixel-Based Fast Direction Detections for H.264 Intra Prediction. IEEE Transactions on Circuits and Systems for Video Technology 18(7), 975–982 (2008)
9. Bjontegaard, G.: Calculation of Average PSNR Differences Between RD-curves. In: The 13th VCEG-M33 Meeting, Austin, TX (2001)

Perceptual Motivated Coding Strategy for Quality Consistency

Like Yu^{1,2}, Feng Dai¹, Yongdong Zhang¹, and Shouxun Lin¹

¹ Institute of Computing Technology, Chinese Academy of Sciences,
100190 Beijing, China

² Graduate University of Chinese Academy of Sciences,
100190 Beijing, China
`{yulike, fdai, zhyd, sxlin}@ict.ac.cn`

Abstract. In this paper, we propose a novel quality control scheme which aims to keep quality consistency within a frame. Quality consistency is an important requirement in video coding. However, many existing schemes usually consider the quality consistency as the quantization parameter (QP) consistency. Moreover, the most frequently used metric to evaluate the quality consistency is PSNR, which has been well known that it is not good for subjective quality evaluation. These flaws of the existing methods are pointed out and proved to be unreasonable. For optimization, we take the effect of texture complexity on subjective evaluation into consideration to build a new D-Q model. We use the new model to adjust the quantization parameters of different regions to keep quality consistency. The simulation result shows that the new scheme gets better subjective quality and higher coding efficiency compared to traditional way.

Keywords: quality consistency, quality fluctuation, video coding, H.264/AVC, perceptual quality.

1 Introduction

Quality control is one of the most important requirements in video coding. In most cases, we demand low quality fluctuation within a frame and between two successive frames [1] because it plays a major negative role on visual perception.

Many schemes have been proposed to achieve constant quality [2-4]. They usually pay attention to the quality fluctuation between frames. The quality difference within a frame is less concerned. The purpose of fluctuation limitation is to provide better perceptual visual feelings. Therefore not only the quality variance between frames should be limited, a uniform visual quality within a frame is also important.

In most cases, researchers restrict the QP (quantization parameter) variation in order to achieve consistent quality. In [5], the authors assign fixed QP to all macroblocks (MBs) in a frame to maintain consistent picture quality. In JVT-G012 [6], which has been adopted in H.264/AVC reference software, the QP variation between each basic unit is limited in order to constrain quality fluctuation. However, in [7] it has been pointed out that coding with a constant value of Q_s (quantization step) generally does not result in either constant bit-rate or constant perceived quality.

In this paper, we propose a new quality control scheme which partitions a frame into several regions and assigns each of them different quantization steps to decrease quality fluctuation within a frame. Different from the existing methods, we abandon the PSNR quality metric because we demand uniform perceptual quality and PSNR is not a good metric for subjective evaluation among different image content. We adopt a “weighted MSE” metric [8], to help control the quality consistency. Based on the new metric, a new D-Q (Distortion-Quantization) model is established. Then different quantization parameters are assigned to different regions according to the new D-Q model and the quality consistency will be optimized. Moreover, the overall quality can be improved via the reasonable coding resource allocation.

The remainder of this paper is organized as follows. Section 2 presents our algorithm and the implementation steps. The experimental results and discussions are shown in Section 3. Finally, we give conclusion in Section 4.

2 Proposed Method

2.1 Problem Analysis

The most common way which uses a fixed quantizer is based on a simple hypothesis: the D-Q characteristics are nearly the same within a frame. Obviously, it does not fit in most cases because D-Q characteristics are source dependent, or content dependent.

Fig. 1 is an example where a fixed QP is assigned to each macroblock. Picture (a) is the 121th frame of “flower garden” (CIF, 352*288). Picture (b) is the reconstructed image encoded by H.264/AVC JM software with the QP set to 40. Obviously the visual quality of the flower part and house part (inside the dashed line) are different. More specifically, the house part looks worse than the flower part although they use the same quantization parameter.

In this case, it is reasonable to increase the quality of the house, or decrease the quality of the flower. However, this adjusting strategy is based on a subjective observation which is unavailable in an encoding system. Therefore a quality evaluation metric should be employed to detect the quality fluctuation.



Fig. 1. Quality fluctuation with a fixed QP

The most commonly used method in video coding is PSNR. Although easily calculated, it has been found to correlate poorly with subjective quality ratings [9]. In this case, the PSNR of the house part is 24.23db and the flower part has only 22.19db, which means it gives an opposite result compared to the observation conclusion. If we use PSNR as the quality evaluation metric, we might improve the quality of the flower part or decrease the house part. Obviously, this will lead to more serious quality fluctuation, although it may have better result in PSNR.

2.2 Distortion Model

Since PSNR/MSE is not good for subjective quality evaluation, we need a new one to replace it. In [8], an MOSp metric was proposed to give a weighted parameter to MSE in order to get more precise evaluation result. It is based on an obvious theory that visibility of artefacts in highly detailed regions is lower than in low detailed regions. The weighted parameter is based on the texture complexity of image content. The simulation result shows that it correlates better to subjective evaluation results than the original MSE. Considering the simplicity and effectiveness of this algorithm, we adopt it to help control the quality consistency. The metric is defined as:

$$\text{Quality} = 1 - k \cdot \text{MSE} \quad (1)$$

Quality has a value range of [0, 1]. 1 represents a perfect quality and 0 is a worst quality. *MSE* is the mean square error of the region. *k* is related to the edge strength of the region which can be referred as:

$$k = 0.03585 * \exp(-0.02439 * \text{EdgeStrength}) \quad (2)$$

where *EdgeStrength* is the average edge strength of the region. The edge strength of a single pixel is computed by Sobel edge detecting filters [10] as:

$$\text{EdgeStrength}(x, y) = |G_{\text{Horizontal}}(x, y)| + |G_{\text{Vertical}}(x, y)| \quad (3)$$

where *G* is the edge magnitude image and (x, y) is the pixel location.

2.3 New D-Q Model

After the new distortion metric selected, we are able to establish a new D-Q model. Firstly we introduce a classic and famous D-Q model, which is used in a zero-mean independent and identically distributed source [11]. It can be written as:

$$D = \frac{Q^2}{\epsilon} \quad (4)$$

where *D* means “distortion”, which is usually denoted by *MSE*. *Q* is quantization step and *ε* is a source dependent parameter. Then (1) and (4) can be combined into:

$$Q^2 = \epsilon \cdot \text{MSE} = \frac{\epsilon}{k} (1 - \text{Quality}) \quad (5)$$

where *D_{new}* is the new distortion level parameter with a value range of [0, 1]. This new D-Q model can be used to help control the quality fluctuation by allocating different quantization parameters to different regions.

2.4 Region Partition

In this section we will discuss how to make use of the new D-Q model to keep consistent quality within a frame.

The new distortion model is based on the detail complexity of the image. Therefore we divide a picture into different detailed regions, such as highly detailed part and low detailed part. The k-means clustering algorithm [12] is employed, which can partition n observations ($\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$) into k sets ($k < n$) $\mathbf{S} = \{S_1, S_2, \dots, S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\operatorname{argmin}_{\mathbf{S}} \sum_{i=1}^k \sum_{x_j \in S_i} (x_j - \mu_i)^2 \quad (6)$$

where μ_i is the mean of S_i .



Fig. 2. Region partition according to the detail complexity

2.5 Our Proposed Scheme

Our new scheme divides a frame into different regions according to their detail complexity (shown in Fig. 2). According to the new D-Q model, each region will be assigned a proper quantization parameter (usually highly detailed region gets larger QP). With the QP adjustment, the quality difference between each region can be minimized and the quality consistency within a frame will be optimized.

The implementation of our coding scheme can be roughly divided into two steps:

Step 1: Calculate the average *EdgeStrength* of each macroblock by Sobel edge detecting filters (3). Then partition them into k sets, and calculate the average *EdgeStrength* of each set, which can be denoted as:

$$\text{Set_AVG_ES}_i = \frac{1}{N_i} \cdot \sum_{j=1}^{N_i} \text{MB_ES}_j \quad (7)$$

where S_i is the i^{th} set, N_i is the MB number of S_i . Then the k values of each set can be achieved by using (2), which is:

$$k_i = 0.03585 * \exp(-0.02439 * \text{Set_AVG_ES}_i) \quad (8)$$

Step 2: A target distortion value has to be set before a frame is encoded, and each region of the current frame will be encoded to approach the target quality. This target value can be a constant one, which will keep the quality of the whole sequence consistent. In another way, the target value can also be adapted based on the quality of the previous frame. The range of the adaption is related to the bit budget remains. Once the target distortion is assigned, quantization steps of each region can be calculated by (5), which can be referred as:

$$Q_i = \sqrt{\frac{\varepsilon_i}{k_i} (1 - D_{T_{\text{target}}})} \quad (9)$$

where k_i can be achieved by (8) and ε_i is source dependent which can be predicted by former encoded frames.

3 Experimental Results and Analysis

Three sequences are chosen for experiments which have significant variation in image complexity. They are “flower_garden”, “coastguard” and “stefan” with the resolution of 352*288. The experiments are based on H.264/AVC reference codec model JM12.2. For each sequence, each frame is divided into two parts, the high detailed and low detailed, using the k-means clustering algorithm.

As mentioned before, our new scheme is designed to decrease visual quality fluctuation within a frame and this will help improve the overall quality. And also we discussed the flaws of PSNR when doing evaluations of different image content in section 2.1. Therefore we will not simply using PSNR comparison to judge the effectiveness of our proposed scheme. The discussion about the coding efficiency is divided into two parts, the objective evaluation and the subjective one.

The objective test is done by encoding the selected sequences with different strategies. Different from traditional tests, the SSIM [13] metric, which is a famous perceptual metric, is employed together with the PSNR for comparison. The results

Table 1. Comparison of coding efficiency

	<i>Ours</i>		<i>JM12.2</i>		<i>Bit-rate (kbps)</i>
	PSNR	SSIM	PSNR	SSIM	
Coastguard	29.35	0.76	29.65	0.724	230
	29.94	0.789	30.3	0.751	290
	30.47	0.811	30.82	0.771	340
	31.31	0.839	31.61	0.802	420
	32.35	0.871	32.62	0.835	560
Flower Garden	25.29	0.648	25.33	0.624	230
	25.95	0.666	26.06	0.642	280
	26.52	0.687	26.61	0.649	310
	27.01	0.695	27.12	0.659	350
Stefan	27.32	0.824	27.56	0.807	210
	28.01	0.835	28.27	0.818	240
	28.66	0.846	28.89	0.831	270
	29.73	0.867	29.99	0.844	330

are shown in Table 1. We can see that our scheme achieves lower PSNR (0.2-0.3db) than JM12.2, which is predictable considering the strategy we carry out. However in SSIM metric, our scheme gets better results. That means our scheme gets better perceptual visual quality.

In order to further prove the efficiency of the new scheme, more subjective tests were done by following the guidelines in ITU-BT.500 [14], involving 12 naive evaluators. The result is shown in Figure 3. Our scheme achieves better subjective quality than the original JM12.2. Two obvious features should be mentioned. Firstly, the quality improvement is larger in lower bandwidth than in higher bandwidth. Secondly, in scenes with large range of texture variation (like flower_garden), the quality enhancement tends to be larger.

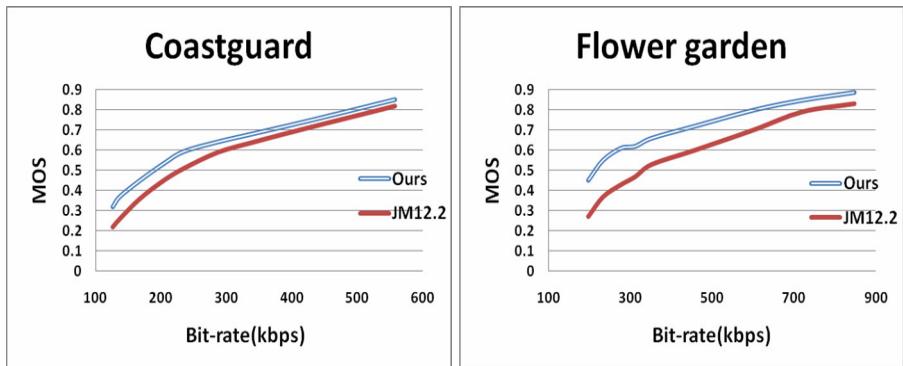


Fig. 3. Results of subjective tests

Figure 4 is a more detailed subjective quality comparison. It is about the 90th frame of “flower garden”. (a) is the original frame. (b) and (c) are encoded by our scheme and JM12.2 respectively, with 311kbps. When evaluated by PSNR metric, (b) and (c) have similar values (26.08db and 26.14db). However, when using SSIM, it seems to have different result. The SSIM metric judges (b) as 0.691 and (c) as 0.646. That means (b) has better quality than (c), which is opposite to the conclusion made by PSNR metric. The subjective evaluation seems to be consistent with conclusions made by SSIM. We can see that the image coded by JM12.2 has poor quality with the tree branch and house roof (too many details are lost). In contrast, our scheme works much better, the whole image looks consistent and has better visual quality.

The reason of the quality improvement is easy to understand. We still take Figure 4 for example. Compared to using fixed QP (JM12.2), our scheme slightly raise the QP of the flower part to save bits which can be used to improve the quality of low detailed regions (such as house and trees). The degradation of flower part is less sensitive because of its high detailed texture, but the improvement of house and trees is much more sensitive. It may cause overall PSNR degradation, but it makes the overall perceptual quality much better. Based on the reasons mentioned above, it is easy to understand that our scheme works better in low bandwidth and in scenes with large range of texture variation.

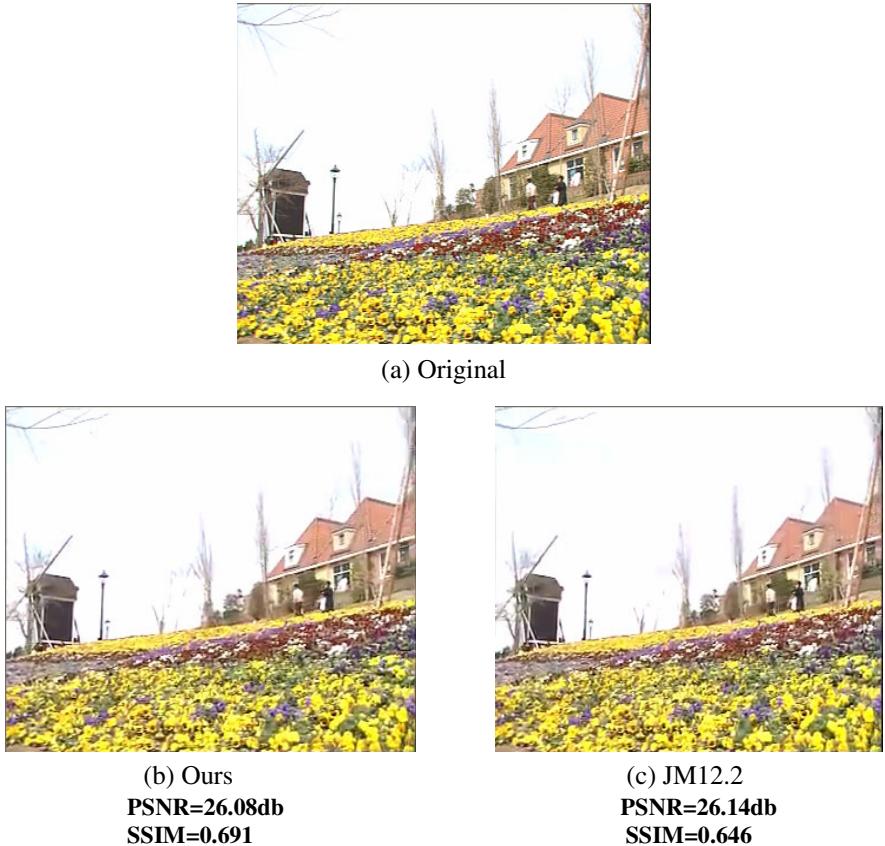


Fig. 4. Subjective evaluation between our scheme and JM12.2 (the 90th frame of “flower garden” with bit-rate at 311kbps)

The extra calculation burden is added into our proposed scheme such as edge strength calculation and clustering in each frame. Therefore the complexity of the additional calculation is tested. Our platform is Intel Q9550 CPU (with 3G RAM), and the operating system is Windows XP. In the case of flower_garden (CIF) sequence, the processing speed of our scheme and original one are 1.67fps and 1.68fps, respectively. Therefore the edge strength and clustering calculations will not affect coding speed much (0.01fps). Our scheme can be used in real-time systems.

4 Conclusions

We have proposed a new quality control scheme which aims to keep quality consistency within a frame. Our new scheme modifies the traditional MSE into texture weighted distortion model. The simulation results show that the new scheme achieves better subjective quality and better coding efficiency, especially in scenes with large range of texture variation. In future work, we plan to implement the coding scheme to smaller control unit, such as MB level, in order to improve the preciseness of quality control.

Acknowledgements

This paper is supported by National Basic Research Program of China (973 Program, 2007CB311100), National Nature Science Foundation of China (60802028), Beijing New Star Project on Science & Technology (2007B071), Co-building Program of Beijing Municipal Education Commission.

References

1. Viterbi, A., Omura, J.: Principles of Digital Communication and Coding. McGraw-Hill, New York (1979)
2. Schuster, G.M., Katsaggelos, A.K.: Rate-Distortion Based Video Compression. Kluwer Academic Publishers, Norwell (1997)
3. Lin, L.J., Ortega, A.: Bit-rate control using piecewise approximated rate-distortion characteristics. *IEEE Trans. Circuits and Systems for Video Technology* 8, 446–459 (1998)
4. Chen, Z., Ngan, K.N.: Distortion variation minimization in real-time video coding. *Signal Processing-Image Communication* 21, 273–279 (2006)
5. Hong, S.H., Yoo, S.J., Lee, S.W., Kang, H.S., Hong, S.Y.: Rate control of MPEG video for consistent picture quality. *IEEE Transactions on Broadcasting* 49, 1–13 (2003)
6. Li, Z., Pan, F., Lim, K.P., Feng, G., Lin, X., Rahardja, S.: Adaptive basic unit layer rate control for JVT. *JVT-G012* (2003)
7. Hoang, D.T., Linzer, E., Vitter, J.S.: Lexicographic bit allocation for MPEG video. *Journal of Visual Communication and Image Representation* 8, 384–404 (1997)
8. Bhat, A., Richardson, I., Kannangara, S.: A new perceptual quality metric for compressed video. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 933–936 (2009)
9. Girod, B.: What's wrong with mean-squared error. In: *Digital Images and Human Vision*. MIT Press, Cambridge (1993)
10. Ran, X., Farvardin, N.: A perceptually motivated three-component image model – Part 1: Description of the model. *IEEE Trans. on Image Processing* 4, 401–415 (1995)
11. Jayant, N., Noll, P.: *Digital Coding of Waveforms*, Englewood Cliffs, NJ (1994)
12. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297 (1967)
13. Wang, Z., Bovik, A.C.: A universal image quality index. *IEEE Signal Processing Letters* 9, 81–84 (2002)
14. ITU-R BT.500 Methodology for the Subjective Assessment of the Quality for TV Pictures, ITU-R Std. (2002)

Compressed-Domain Shot Boundary Detection for H.264/AVC Using Intra Partitioning Maps

Sarah De Bruyne, Jan De Cock, Chris Poppe, Charles-Frederik Hollemeersch,
Peter Lambert, and Rik Van de Walle

Ghent University - IBBT,

Department of Electronics and Information Systems - Multimedia Lab

Gaston Crommenlaan 8 bus 201, B-9050 Ledeberg-Ghent, Belgium

sarah.debruyne@ugent.be

<http://multimedialab.elis.ugent.be>

Abstract. In this paper, a novel technique for shot boundary detection operating on H.264/AVC-compressed sequences is presented. Due to new and improved coding tools in H.264/AVC, the characteristics of the obtained sequences differ from former video coding standards. Although several algorithms working on this new standard are already proposed, the presence of IDR frames can still lead to a low accuracy for abrupt transitions. To solve this issue, we present the motion-compensated intra partitioning map which relies on the intra partitioning modes and the motion vectors present in the compressed video stream. Experimental results show that this motion-compensated map achieves a high accuracy and exceeds related work.

Keywords: Shot boundary detection, video analysis, compressed domain, H.264/AVC.

1 Introduction

During the last decades, a significant increase in the use and availability of digital multimedia content can be witnessed. Unfortunately, these video collections often lack information related to the structure and the actual content of the video. When accessing these video streams in case no metadata is available, time-consuming, sequential scanning is the only option. As a consequence, to facilitate multimedia consumption, intensive research has been done in the domain of indexing, retrieval, browsing, and summarization. Since the identification of the temporal structure of video is an essential task for many video indexing and retrieval applications, the first step commonly taken for video analysis is shot boundary detection as shots are the basic units for a large majority of video content analysis algorithms [5]. According to whether the transition between consecutive shots is abrupt or not, boundaries are classified as cuts or gradual transitions.

In order to preserve storage space and to reduce bandwidth constraints, most video data is available in compressed form. By relying on compressed-domain features which can be extracted directly from the compressed bitstream,

time-consuming decompression can be avoided and coarse but potentially useful information present in the bitstream can efficiently be reused. Consequently, compressed-domain algorithms for shot boundary detection are gaining importance. In the past, many compressed-domain algorithms were proposed which rely on the MPEG-1 Video and MPEG-2 Video standards. However, as the H.264/AVC video coding standard [12] performs significantly better than any prior standard in terms of coding efficiency, more video content will be coded in this video format in the future. Its superior compression performance can mainly be attributed to the new or improved coding tools. However, these coding tools influence the compressed domain features to a great extent and render prior algorithms working on MPEG-1 Video and MPEG-2 Video obsolete. As a consequence, recently, efforts have been undertaken to design new shot boundary detection algorithms working on H.264/AVC.

The outline of this paper is as follows. Section 2 addresses related work and remaining issues in the area of shot boundary detection algorithms operating on H.264/AVC-compressed video streams. Section 3 introduces our novel algorithm to detect shot boundaries, whereas results are provided in Section 4. Conclusions are drawn in Section 5.

2 Related Work

2.1 General Techniques

In literature, most of the algorithms work on MPEG-1 Video and MPEG-2 Video. On the one hand, DCT coefficients are exploited. Arman *et al.* [1] compare a subset of DCT coefficients of two successive I frames to calculate the frame differences as the coefficients in the frequency domain are mathematically related to the spatial domain. The temporal resolution of this type of techniques is low, resulting in an increased amount of false alarms when camera motion is present. Furthermore, Yeo and Liu defined the concept of DC images [13], which are spatially reduced versions of the original image and which are generated by only taking into account the first DCT coefficient in each block, i.e., the DC coefficient, and motion vectors. Based on these DC images, similarity metrics defined for color features in the pixel domain can be modified to operate in the compressed domain.

On the other hand, the distribution of the different macroblock types and motion information [4][11] can also be used as features to detect shot boundaries. When an abrupt transition occurs between two successive P pictures, it is expected that a significant amount of macroblocks in the second frame is intra coded since these macroblocks cannot be predicted well from prior reference frames. The prediction directions of motion vectors in intermediate B frames can then be utilized to detect the exact location of the transition.

Although most algorithms mainly focus on the detection of abrupt transitions, the aforementioned features can also be used to detect gradual transitions [2]. Due to the large variety in terms of effects and duration, the accuracy for gradual transitions is typically inferior to the abrupt transitions.

2.2 Algorithms for H.264/AVC

H.264/AVC contains a number of new or improved coding tools which have a major impact on the aforementioned shot boundary detection algorithms. Firstly, intra prediction in H.264/AVC is conducted in the spatial domain by relying on neighboring samples of previously-decoded blocks in order to reduce the spatial redundancy in images. As such, DC coefficients in intra-coded pictures no longer represent average energy, but only represent an energy difference. Consequently, shot boundary detection algorithms working on DC images can no longer be applied to H.264/AVC bitstreams. The second feature, multiple reference picture motion compensation, allows an encoder to not only use the previous and following reference frame during encoding, but makes it possible to also use additional priorly-decoded frames as reference. As this results in vagueness about random access in the bitstream, the concept of Instantaneous Decoding Refresh (IDR) was introduced [12]. This special I frame indicates that no subsequent pictures in the bitstream will require references to pictures prior to the IDR picture in decoding order. The prediction chain is broken; hence it is insufficient to only rely on reference directions to detect shot boundaries.

Up to now, a few algorithms working on H.264/AVC-compressed bitstreams have been published. In [8], Kim *et al.* define a dissimilarity metric for I frames by relying on macroblock partitions (i.e., Intra $_4 \times 4$ and Intra $_{16 \times 16}$ prediction). A more complex discontinuity metric based on solely I frames is proposed by Kuo and Lo where intra mode histogram distances are calculated based on the different intra prediction mode directions [9]. However, the exact location of a shot boundary cannot be determined by these two algorithms as information from P and B frames is not taken into consideration. Extensions on this second algorithm are proposed in [10] and [14] to deal with all type of frames. Firstly, these algorithms exploit the intra prediction histograms to locate potential groups of pictures (GOPs) where the probability of a shot boundary is high. Secondly, the different inter prediction modes and motion vector directions of the intermediate P and B frames are used to locate the exact location of the transition by making use of thresholds or Hidden Markov Models. However, due to the presence of IDR frames, shot boundaries located just before the IDR frames are falsely detected in case the dissimilarity between the two I frames is relatively large. Furthermore, when the dissimilarity metric is low but both I frames actually belong to different shots, the temporal prediction chain is not considered, leading to missed detections.

In our previous work [3], these two problems were tackled. To locate the abrupt transitions, the temporal dependencies between successive frames are examined first. Secondly, when encountering an IDR frame breaking this prediction chain, spatial dissimilarities are considered. In contrast to the aforementioned related algorithms, the changing characteristics of the content are taken into account by constructing a “static intra partitioning map”. In particular, the static intra partitioning map is updated when new intra-coded macroblocks residing at intermediate, inter-coded frames are encountered. Gradual changes are detected by relying on the percentage of intra-coded macroblocks. Since fast global and

local motion can result in similar patterns as gradual transitions, motion-activity intensity is considered as well to identify the exact origin of the content change.

Although the static intra partitioning map already solves several issues, its accuracy is still inferior to video streams compressed without IDR frames. When examining the origin of the false alarms, it can be seen that they mostly occur when the content slightly changes as a result of motion, which is generally compensated by using inter prediction. In the next section, we will extend the algorithm proposed in [3] by introducing the motion-compensated intra partition map which overcomes this problem by relying on motion compensation techniques.

3 Intra Partitioning Maps

Whereas the abrupt transitions located in between two inter-coded frames can be detected by analyzing the main reference directions of the frames, another technique is required to determine whether a shot boundary is present in between an IDR frame and its preceding frames. To solve the issue of the broken temporal prediction chain, algorithms designed for prior video coding standards would typically make use of DC images to compare the spatial characteristics of successive frames. However, due to the introduction of spatial intra prediction in H.264/AVC, these iconic versions of the content can no longer be generated without further decoding the bitstream. Therefore, we employ intra partitioning maps which reflects the spatial dissimilarities between frames based on the selected intra partitioning modes.

H.264/AVC supports multiple intra macroblock partitions, i.e., Intra_4×4, Intra_8×8 and Intra_16×16. The first mode is generally selected by an encoder in case of significant detail, whereas the Intra_16×16 mode is preferred for smooth areas. When coding high-resolution video sequences using the High profile of H.264/AVC, Intra_8×8 will also be selected and mainly replace Intra_4×4 modes to code the high-textured areas. As the subdivision in different macroblock partitions roughly reflects the detail of the content, comparing the distribution of two frames can be used to estimate the spatial dissimilarities, as illustrated in Fig. II.

Comparing the current I frame with the previous I frame is not recommended as the content can change significantly between these two intra-coded frames. For example, a shot boundary can be located at an intermediate P or B frame, new objects can appear, or camera motion can occur (Fig. II). Therefore, we introduce an *intra partitioning map* M_i indicating which intra partitioning modes most likely correspond to the content of the intermediate, inter-predicted frames. This partitioning map can be constructed in different ways. In [3], the static intra partitioning map was introduced, which considers all intra-coded macroblocks in intra- as well as inter-coded frames. In this paper, the motion-compensated intra partitioning map is introduced, which extends the static map by including motion information.

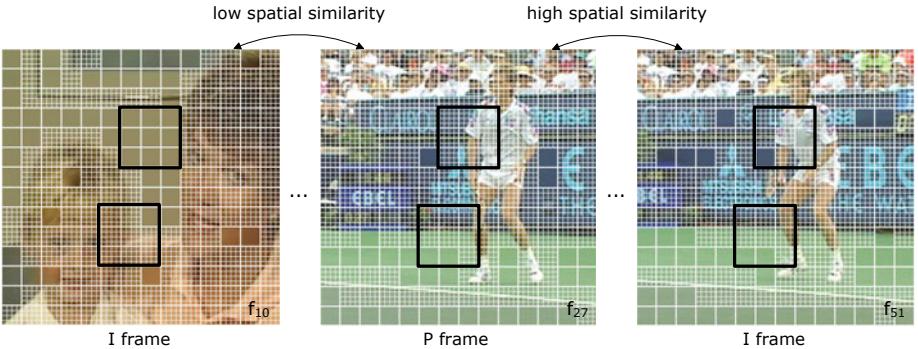


Fig. 1. Distribution of Intra₄×4 and Intra₁₆×16 macroblocks. Although the second frame is a P frame, it is mainly intra coded as it is the first frame of a new shot. As such, it is important to update the intra partitioning maps with information from inter-coded frames.

Static intra partitioning map. A first approach to construct an intra partitioning map for the inter-coded frame f_i is by remembering for each macroblock position the intra partitioning mode which was last encountered. To put it differently, the intra partitioning map M_i of frame f_i is constructed by updating the previous map M_{i-1} with the partitioning modes of the intra-coded macroblocks in the current frame. As such, this map can be used to represent the spatial distribution of the content of the current frame, in spite of the fact that this frame mainly contains inter-coded macroblocks. By comparing the current I frame f_i with the static map M_{i-1} , the spatial dissimilarity between the current and previous frame can be calculated. However, instead of comparing partitioning modes at corresponding positions, a window of several macroblocks is selected for each macroblock. This way, small movement of objects or the camera will lead to fewer false alarms.

Motion-compensated intra partitioning map. The downside of the static intra partitioning map is its limited support for motion. In particular, when the difference between two I frames belonging to the same shot is large, resulting from the large distance between these two frames or from relatively fast moving objects or the camera, this window will not be able to cover the displacement of the scene. As a result, the amount of falsely detected shot boundaries will increase. To overcome this problem, motion information can be considered in order to construct a motion-compensated intra partitioning map.

Instead of copying the partitioning modes of the previous intra partitioning map, we propose to use the partitioning modes of the reference blocks to which the motion vectors (MVs) of the current frame point to. The favorable aspect of this motion compensation step is shown in Fig. 2(a) and 2(c) and further explained below. The macroblock in f_{186} which is marked in red is part of the low-textured sky (Fig. 2(c)) and would typically result in an Intra₁₆×16 partitioning mode when intra prediction would be applied, as verified by coding the sequence with I frames only. When copying the partitioning mode of the



Fig. 2. Between the two reference frames f_{184} and f_{186} of the Foreman sequence, the camera is panning to the right side. To update the motion-compensated intra partitioning map M_{186} , the MVs of inter-coded macroblocks in f_{186} are used to locate the corresponding intra partitioning modes in f_{184} (a and c). The intra-coded macroblocks in f_{186} are used to update the intra partitioning modes (d).

macroblock located at the same location in the previous partitioning map, marked by the semi-transparent red rectangle in Fig. 2(a), incorrect spatial information belonging to the high-textured crane would be passed on. However, thanks to the motion compensation step, correct spatial information can now be stored in the partitioning map, which corresponds to the block indicated by the full red rectangle in Fig. 2(a). Obviously, for intra-coded macroblocks, the new partitioning modes are used to update the partitioning map (Fig. 2(d)).

The reference block to which the MV of the current frame points to does not necessarily coincide with macroblock or sub-macroblock boundaries. As such, this reference block can overlap with different partitioning modes, as illustrated by the green block in Fig. 2(a). Therefore, it is undesired to store only one partitioning mode for each block. Instead, we propose to store the likelihood of each possible partitioning mode. As multiple partitioning modes for inter-coded macroblocks

exist and each partition contains its own MVs, it is desired to divide the intra partitioning map into a uniform field of basic units of 4×4 pixels, which corresponds to the smallest partition for which MVs can change. This subdivision also results in a finer granularity, improving the accuracy of the likelihoods of each partitioning mode.

In order to calculate the likelihoods of the different partitioning modes for an inter-coded block b in the motion-compensated intra partitioning map M_i , i.e., $M_i[b, mode]$, the likelihoods of the blocks in the reference map M_{ref} which overlap with the reference block are considered. To incorporate the percentage of overlap between the motion-compensated reference block and the overlapping block r , a new variable $OverlappingSize_r$ is introduced. This leads to the following formula where $OverlappingBlocks_b$ is defined as the set of overlapping blocks connected to block b :

$$M_i[b, mode] = \sum_{r \in OverlappingBlocks_b} OverlappingSize_r \cdot M_{ref}[r, mode]. \quad (1)$$

When an intra-coded macroblock is encountered, the sixteen corresponding blocks in the intra partitioning map are set to one for the encountered partitioning mode, whereas the other modes are set to zero.

Let f_i denote the current I frame and M_i the intra partitioning map containing the new intra partitioning modes, M_{i-1} the map of the previous frame, and B_4 the amount of 4×4 blocks in a frame. Define $Modes$ as the set of possible macroblock partitions ($Modes = \{Intra_4 \times 4, Intra_8 \times 8, Intra_16 \times 16\}$). Furthermore, let M be a placeholder for M_i and M_{i-1} , and n and m macroblocks. The dissimilarity metric Ω used to compare the current I frame f_i and the preceding motion-compensated intra partitioning map can then be defined by the average of the dissimilarity values of all blocks b in f_i .

$$\Omega(f_i) = \frac{1}{B_4} \sum_{b \in f_i} \omega(f_i, b). \quad (2)$$

This block dissimilarity value $\omega(f_i, b)$ is obtained by comparing a window around the current block b in the intra partitioning map M_i with the collocated window in M_{i-1} . This comparison is done by making histograms of the various partitioning modes present in both windows, and thereafter, calculating the difference between these histograms.

$$\omega(f_i, b) = \frac{\sum_{t \in Modes} |s_{b,t}^{M_i} - s_{b,t}^{M_{ref}}|}{2 \cdot window\ size}. \quad (3)$$

For this purpose, the percentages of all blocks which are coded using a certain coding partitioning mode t located in the window around block b in partitioning map M_i or M_{i-1} are added up (i.e., $s_{b,t}^M$).

$$s_{b,t}^M = \sum_{n \in window(b)} M[n, t]. \quad (4)$$

Threshold selection. To decide whether a shot boundary is located in between the current I frame f_i and the preceding frame f_{i-1} , the spatial dissimilarity $\Omega(f_i)$ needs to be compared to a threshold T . Although predefined, static thresholds which remain the same over the entire sequence are often applied, this type of thresholds cannot be adjusted to local properties of the sequences. As the obtained results do not represent probabilities, but rather indicate the difference compared to previous frames, we prefer to work with an adaptive threshold.

To adapt T to the local properties of the content, the M previous spatial dissimilarity values calculated for I frames are considered. Statistical information such as the mean and variation obtained from these M elements is then used to determine the local, content-dependent threshold T . Let μ_Ω denote the mean of the dissimilarity values of the M previous I frames and their preceding intra partitioning map and σ_Ω the corresponding standard deviation. T_{intra} can then be defined as:

$$T = \mu_\Omega + \alpha\sigma_\Omega. \quad (5)$$

The values for M and α are computed heuristically, resulting in typical values lying around 8 and 6 respectively. Furthermore, a lower boundary is set to this threshold to avoid extreme values. In particular, when the start of a shot is static, the corresponding dissimilarity values are close to zero. A small amount of motion further in the shot would otherwise result in a false alarm. When observing typical values for $\Omega(i)$ corresponding to abrupt transitions, an appropriate lower boundary is 0.2.

4 Performance Results

Several video sequences with various characteristics in terms of resolution, length, quality, and content were selected to evaluate the performance of our shot boundary detection algorithm. The first two sequences originate from the publicly available MPEG-7 Content Set [6] and represent a part of a news sequence and a basketball sequence. Due to the low quality and resolution, three more recent, proprietary sequences were added to the test set as well.

These sequences were coded using the Joint Model reference software and Main profile enabled, which is suitable for temporal-segmentation applications. IBP GOP structures and a quantization parameter (QP) of 26 were selected. Furthermore, two configurations were made: once with I frames and once with IDR frames, which were inserted every 32 frames. As such, the effect of intra partitioning maps can be evaluated when the prediction chain is broken (i.e., when using IDR frames). As the static and motion-compensated intra partitioning maps only influence the detection of abrupt transitions, the accuracy results for gradual transitions are discarded from the accuracy results presented in Table II.

When comparing the algorithms working on I-frame and IDR-frame sequences, a small decrease in precision is observed for some sequences, which can be attributed to the gaps in the temporal prediction chain resulting from IDR frames, which are falsely marked as abrupt transition. In general, the results of the motion-compensated intra partitioning map exceed the static intra partitioning

Table 1. Accuracy in precision (P) and recall (R) (%) of the motion-compensated (MC) intra partitioning map. For comparison, the results for the static intra partitioning map as well as for the sequences coded using classic I frames are depicted. Furthermore, the results for the “T2I Shotdetection TrecVID 2004” algorithm are provided as reference.

Test sequence	I frames		IDR frames static map		IDR frames MC map		T2I algorithm	
	P	R	P	R	P	R	P	R
News 1	91	100	93	99	95	99	93	97
Basket	94	98	94	98	94	98	97	94
News 2	98	100	91	99	98	99	91	99
Soap	99	100	95	99	99	99	99	91
Trailer	100	100	100	100	100	100	95	99

Table 2. Influence of the quantization parameter on the accuracy of the intra partitioning map for the News 2 sequence

QP	I frames		IDR frames static map		IDR frames MC map	
	P	R	P	R	P	R
26	98	100	91	99	98	99
30	99	99	88	99	96	99
34	100	99	91	99	97	99
38	100	99	93	98	99	98
42	99	97	97	96	98	96

map. This difference can mainly be attributed to shots containing medium object or camera motion. As the content between successive I frames clearly changes but the inter-coded prediction still obtains good results, the static map will not be updated whereas the motion-compensated map better reflects the content change.

The effect of the selected QP should not be neglected as this influences the selected intra partitioning modes and the amount of intra-coded macroblocks. As shown in Table 2, it catches the eye that the detection of abrupt transitions in I-frame sequences is hardly affected by the different QP values. For the IDR-frame sequences analyzed using the static and the motion-compensated intra partitioning map, the recall values in average slightly decrease. Due to the reduced amount of intra-coded macroblocks when applying higher QP values, the partitioning map is updated less frequently with intra-coded macroblock information and therefore is less accurate. As a result, the variable threshold will typically obtain higher values, resulting in more missed detections.

The interpretation of the obtained accuracy results can be enhanced by comparing these results with related work. Therefore, the results of the publicly available “T2I Shotdetection TrecVID 2004” algorithm [7] that works in the pixel domain are added to Table 1. To detect abrupt transition candidates, this pixel-domain approach uses color histogram values which are calculated within

a five frames width window and the edge change ratio calculated between consecutive frames as well as frames at a distance of ten. To confirm or reject the candidates, a block-based motion analysis is further used. On the one hand, the missed detections are primarily caused by succeeding shots covering one scene, leading to very similar color histograms. On the other hand, the origin of the false alarms are lighting changes and camera motion, even though block-based motion analysis already performed to reject some candidates.

When comparing the results of the uncompressed-domain algorithms with the results of the proposed algorithm, it can be seen that our proposed algorithm can compete in terms of accuracy. Although pixel-domain algorithms can rely on more features, they often do not perform complex calculations in order to limit the execution time.

5 Conclusions

In this paper, we have introduced the concept of motion-compensated intra partitioning maps, which are used to automatically detect shot boundaries in H.264/AVC sequences. Due to the introduction of IDR frames, it is insufficient to only investigate temporal dependencies. Consequently, spatial features need to be considered to reduce the false alarms. In this paper, the motion-compensated intra partitioning map is proposed which aims at correctly detection shot boundaries near IDR frames. Experimental results show that the proposed motion-compensated intra partitioning map obtains a higher accuracy compared to the static intra partitioning map as this latter technique experiences more problems when medium object or camera motion is present.

Acknowledgments

The research activities that have been described in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT-Flanders), and the Fund for Scientific Research-Flanders (FWO-Flanders).

References

1. Arman, F., Depommier, R., Hsu, A., Chiu, M.Y.: Content-based browsing of video sequences. In: Proceedings of ACM Multimedia, pp. 97–103 (1994)
2. Bescós, J.: Real-time shot change detection over online MPEG-2 video. IEEE Transactions on Circuits and Systems for Video Technology 14(4), 475–484 (2004)
3. De Bruyne, S., Van Deursen, D., De Cock, J., De Neve, W., Lambert, P., Van de Walle, R.: A compressed-domain approach for shot boundary detection on H.264/AVC bit streams. Signal Processing: Image Communication - Semantic Analysis for Interactive Multimedia Services 23(7), 473–498 (2008)

4. Fernando, W.A.C.: Sudden scene change detection in compressed video using interpolated macroblocks in B-frames. *Multimedia Tools and Applications* 28(3), 301–320 (2006)
5. Hanjalic, A.: Towards theoretical performance limits of video parsing. *IEEE Transactions on Circuits and Systems for Video Technology* 17(3), 261–272 (2007)
6. ISO/IEC: Description of MPEG-7 content set. ISO/IEC JTC1/SC29/WG11/N2467 (October 1998)
7. Jacobs, A., Miene, A., Ioannidis, G.T., Herzog, O.: Automatic shot boundary detection combining color, edge, and motion features of adjacent frames. In: TRECVID 2004 Workshop Notebook Papers, pp. 197–206 (2004)
8. Kim, S.-M., Byun, J.W., Won, C.S.: A scene change detection in H.264/AVC compression domain. In: Ho, Y.-S., Kim, H.-J. (eds.) *PCM 2005*. LNCS, vol. 3768, pp. 1072–1082. Springer, Heidelberg (2005)
9. Kuo, T.Y., Lo, Y.C.: Detection of H.264 shot change using intra predicted direction. In: Proceedings of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 204–207 (2008)
10. Liu, Y., Wang, W., Gao, W., Zeng, W.: A novel compressed domain shot segmentation algorithm on H.264/AVC. In: Proceedings of the IEEE International Conference on Image Processing, vol. 4, pp. 2235–2238 (2004)
11. Pei, S.C., Chou, Y.Z.: Efficient MPEG compressed video analysis using macroblock type information. *IEEE Transactions on Multimedia* 1(4), 321–333 (1999)
12. Wiegand, T., Sullivan, G.J., Bjøntegaard, G., Luthra, A.: Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology* 13(7), 560–576 (2003)
13. Yeo, B.L., Liu, B.: Rapid scene analysis on compressed video. *IEEE Transactions on Circuits and Systems for Video Technology* 5(6), 533–544 (1995)
14. Zeng, W., Gao, W.: Shot change detection on H.264/AVC compressed video. In: Proceedings of the IEEE International Symposium on Circuits and Systems, vol. 4, pp. 3459–3462 (2005)

Adaptive Orthogonal Transform for Motion Compensation Residual in Video Compression

Zhouye Gu, Weisi Lin, Bu-sung Lee, and Chiew Tong Lau

School of Computer Engineering,
Nanyang Technological University,
Singapore, 639798

{guzh0001,wslin,ebslee,asctlau}@ntu.edu.sg

Abstract. Among the orthogonal transforms used in video and image compression, the Discrete-Cosine-Transform (DCT) is the most commonly used one. In the existing video codecs, the motion-compensation residual (MC-residual) is transformed with the DCT. In this paper, we propose an adaptive orthogonal transform that performs better on the MC-residual than the DCT. We formulate the proposed new transform based on L1-Norm minimization with orthogonal constraints. With the DCT matrix as the starting point, it is guaranteed to derive a better orthogonal transform matrix in terms of L1-Norm minimization. The experimental results confirm that, with little side information, our method leads to higher compression efficiency for the MC-residual. Remarkably, the proposed transform performs better in the high/complex motion situation.

Keywords: Motion Compensation Residual, Transform Coding, L1-Norm Minimization, Orthogonal Constraints, Video Coding.

1 Introduction

Transformation is an essential part of digital video and image compression. In video coding, motion-compensation residual (MC-residual) is transformed for compression. Usually, motion compensation works well in smooth and slowly moving regions, and the MC-residual in such regions is small. However, for the high/complex motion areas, motion compensation doesn't work well. In existing video codecs, the original image and the MC-residual image are both transformed by DCT. However, the spatial characteristics of the MC-residual image are different from that of an original image. A number of studies reported that statistically, pixels in the MC-residual tend to have much less correlation compared with those of nature images [1], [2], [3].

The conversion of spatial domain pixels into frequency coefficients is efficient when the pixels are strongly correlated. As shown in Figure 1(a) for an image, most of the energy after DCT is concentrated with the top-left coefficients, which is highlighted in the red box. However, there is little correlation between pixels in MC-residual signals, and therefore high magnitude transform coefficients are distributed among low and high frequency areas, as shown in Figure 1(b); in this case, DCT is less efficient for data compression purpose.

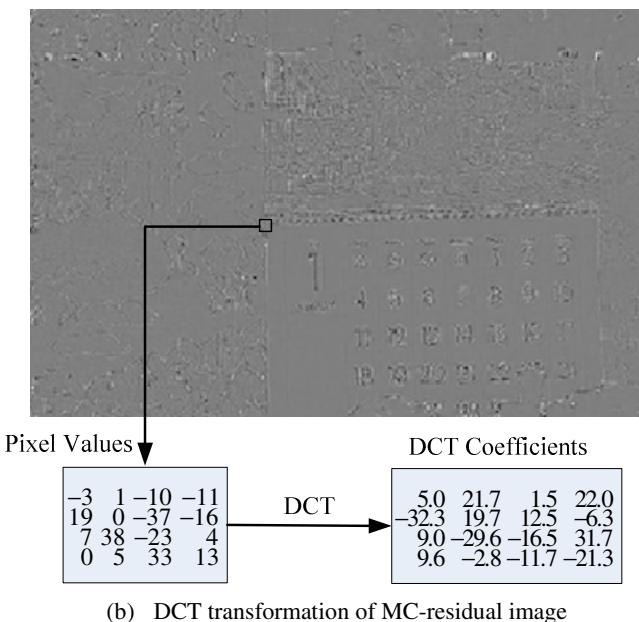
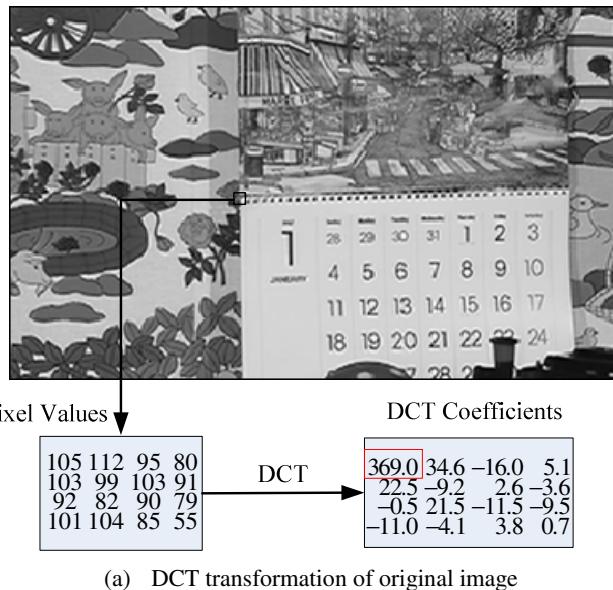


Fig. 1. DCT transformation result of original image and MC-residual image

Other transforms, such as Discrete Wavelet Transform (DWT), Singular Value Decomposition (SVD), Karhunen-Loëve transform (KLT) and their combinations [4],[5],[6] have also been explored for better coding efficiency based on the local

statistical characteristics in images. In [5], each block is coded by either DCT or SVD based on its correlation. In [6], according to the standard deviation (STD) of each block, either DWT or SVD is applied to code these blocks. For DCT and DWT, the transform is fixed for all images, while for SVD it is adaptive to every image.

There is no specially designed adaptive orthogonal transform for MC-residual, and in this work, we aim at a new orthogonal transformation with which most of the energy of a MC-residual image is compacted with a few coefficients, so that we can achieve efficient data compression; since such compaction is only possible by adaptively defining or selecting a transform according to visual content, keeping side information (for the transform defined/selected) small is the key to the success (otherwise, the benefit of compaction would be off-set by the expense of side information).

In this paper, we propose a new, adaptive orthogonal transform for MC-residual compression based on L1-Norm minimization under orthogonal constraints, with little side information to be sent to the decoding end. The next section first introduces and analyzes the problem we are going to solve. We present the detailed iterative solution in Section 3. Experimental results, presented in Section 4, show that the proposed improves the compression efficiency of the MC-residual. Section 5 concludes this paper.

2 Problem Formulation

We formulate the problem of compression of MC-residual as follows:

$$\begin{aligned} Y_{opt} = \arg \min_Y \sum_i \|Vec(Y^T X_i Y)\|_1 \\ s.t \quad Y^T \bullet Y = I \end{aligned} \quad (1)$$

where X_i is a given n -by- n matrix, which represents the i -th block of a MC-Residual image, Y is a matrix with n -by- n variables and I is a n -by- n identity matrix. The operator $Vec()$ is a vectorization operator, which turns all the matrix elements to a vector form in the row by row order. Suppose A is a n -by- n matrix, after vectorization we have the vector in (2), where each a_{ij} is an element of A at the position (i,j).

$$Vec(A) = [a_{1,1}, a_{1,2} \dots a_{1,n}, a_{2,1}, a_{2,2}, \dots a_{2,n}, \dots a_{n,1}, a_{n,2} \dots a_{n,n}] \quad (2)$$

If Y is the DCT basis, then $Y^T X_i Y$ is the classical 2D-DCT. The L1-Norm minimization problem has the sparse property [7]; however, the objective function

$$f(Y) = \sum_i \|Vec(Y^T X_i Y)\|_1 \quad (3)$$

is not a convex function, while the constraints are also not convex since they are a set of non-linear equations. Sparse coding and L1-Norm optimization problems have recently received much attention for research efforts [7],[8],[9]. In this paper, we deal with the orthogonal constraints by a set of rotation matrices. So we convert the non-convex constraints in (1) into convex ones. As a result, it is guaranteed to achieve a better orthogonal matrix than the given initial matrix in terms of L1-Norm minimization as analysed in the rest of the section.

In (1), $Y \in R^{n \times n}$ and $Y^T Y = I$. The constraints $Y^T Y = I$ contain $n(n+1)/2$ equations; therefore, although Y seems to have n^2 variables, it has only $n(n-1)/2$ independent variables, because $n(n-1)/2 = n^2 - n(n+1)/2$.

Since an orthogonal matrix could be derived by rotating a predefined orthogonal matrix, e.g. the DCT basis, in the R^n space, we represent these $n(n-1)/2$ independent variables with $n(n-1)/2$ rotation degrees $\theta_{1,1}, \dots, \theta_{n(n-1)/2}$. We derive any orthogonal matrix by multiplying a series of $n(n-1)/2$ rotation matrices with a predefined orthogonal matrix. We construct each rotation matrix $R_{i,j}$ as follows. Given a n-by-n identity matrix I , we set

$$\begin{aligned} I(i,i) &= \cos \theta_{i,j} & I(i,j) &= \sin \theta_{i,j} \\ I(j,i) &= -\sin \theta_{i,j} & I(j,j) &= \cos \theta_{i,j} \end{aligned}$$

where $1 \leq i < j \leq n$. In this way, we get rotation matrix $R_{i,j}$.

If $n = 6$, then

$$R_{1,4}(\theta_{1,4}) = \begin{bmatrix} \cos \theta_{1,4} & 0 & 0 & \sin \theta_{1,4} & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ -\sin \theta_{1,4} & 0 & 0 & \cos \theta_{1,4} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

where $\theta_{1,4}$ is a variable.

If we denote the n -by- n DCT basis as matrix Y_{DCT} , then we can get any other orthogonal basis $Y_{orthogonal}$ by

$$Y_{orthogonal}(\boldsymbol{\theta}) = (\prod_{i=1}^{n-1} \prod_{j=i+1}^n R_{i,j}(\theta_{i,j})) * Y_{DCT} \quad (4)$$

where $\boldsymbol{\theta}$ is a $n(n-1)/2$ -by-1 column vector. With this transformation, we eliminate the orthogonal constraints and convert problem (1) into the following problem with non-convex objective function, subject to convex constraints:

$$\begin{aligned} \min f(\boldsymbol{\theta}) &= \sum_i \|Vec(Y_{orthogonal}(\boldsymbol{\theta})^T X_i Y_{orthogonal}(\boldsymbol{\theta}))\|_1 \\ s.t. \quad & -\frac{\pi}{2} \bullet \mathbf{I} \leq \boldsymbol{\theta} < \frac{\pi}{2} \bullet \mathbf{I} \end{aligned} \quad (5)$$

where \mathbf{I} is a $n(n-1)/2$ -by-1 column vector with all the elements equal 1. As for $Y_{orthogonal}$, we set Y_{DCT} as the predefined orthogonal matrix to start with, so the proposed minimization method can derive a better orthogonal transform than DCT.

With this transformation, although we are not guaranteed to get the global minimum, we can search around the predefined orthogonal matrix, e.g. Y_{DCT} , and get the local minimum by a gradient-based method, since now the domain of the variable $\boldsymbol{\theta}$ is convex. More importantly, in this way, we are guaranteed to get a better orthogonal matrix in terms of L1-Norm minimization, unless the predefined matrix is already the local minimum.

To derive the gradient $\nabla \boldsymbol{\theta}$, we resort to the numerical computation, since L1-Norm is not differentiable. We calculate the partial derivative with

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \theta_{i,j}} \approx \frac{f(\boldsymbol{\theta} + \delta \cdot \mathbf{e}_{i,j}) - f(\boldsymbol{\theta} - \delta \cdot \mathbf{e}_{i,j})}{2\delta} \quad (6)$$

Where δ is a very small positive number, and $\mathbf{e}_{i,j}$ is a $n(n-1)/2$ -by-1 vector, with all the elements being equal zeros except the element at the corresponding location of $\theta_{i,j}$ equals 1; we get

$$\nabla \boldsymbol{\theta} = \left(\frac{\partial f(\boldsymbol{\theta})}{\partial \theta_{1,2}}, \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_{1,3}}, \dots, \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_{n-2,n-1}}, \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_{n-2,n}}, \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_{n-1,n}} \right)^T \quad (7)$$

where $\nabla \boldsymbol{\theta}$ is a $n(n-1)/2$ -by-1 vector, which is used for gradient-based search specified in the next section.

3 The Proposed Algorithm

The algorithm used for searching the local minimum around the predefined orthogonal matrix is a gradient-based search described in Fig. 2.

By this algorithm, we can get to at least the local minimum at $\boldsymbol{\theta}_{opt}$. Then the desired optimal orthogonal matrix is $Y_{orthogonal}(\boldsymbol{\theta}_{opt})$. Based on our experiment, by setting $\delta=10^{-4}$, $\beta=0.98$, tolerance $\varepsilon=0.001$ and stepsize $S=1.5$, we can get the local minimum within 10 iterations (the loop counted by i in the algorithm below) for most of the video sequences.

Inputs: All the blocks from a MC-residual image $\{X_1, X_2, \dots, X_n\}$

$$X_i \in R^{nxn}$$

$$\text{The DCT matrix } Y_{DCT} \in R^{nxn}$$

Outputs: The optimal degree vector $\boldsymbol{\theta}_{opt}$

1. Initialization: $\boldsymbol{\theta}^{(1)} = \mathbf{0}$, $0 < \beta < 1$, $i \leftarrow 1$, step size $S > 0$, tolerance ε

2. Calculate $\nabla \boldsymbol{\theta}^{(i)}$ according to Eq. (5),(6)

$$2.1 \quad \boldsymbol{\theta}_1^{(i)} = \boldsymbol{\theta}^{(i)} = \beta \cdot S \cdot \nabla \boldsymbol{\theta}^{(i)} / \|\nabla \boldsymbol{\theta}^{(i)}\|_2, k \leftarrow 2;$$

$$2.2 \quad \boldsymbol{\theta}_k^{(i)} = \boldsymbol{\theta}^{(i)} = \beta^k \cdot S \cdot \nabla \boldsymbol{\theta}^{(i)} / \|\nabla \boldsymbol{\theta}^{(i)}\|_2$$

2.3 If $f(\boldsymbol{\theta}_k^{(i)}) > f(\boldsymbol{\theta}_{k-1}^{(i)}) \& f(\boldsymbol{\theta}_{k-1}^{(i)}) < f(\boldsymbol{\theta}^{(i)})$

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}_{k-1}^{(i)}, \text{ goto Step 3}$$

else $k \leftarrow k+1$, goto Step 2.2

3. If $|f(\boldsymbol{\theta}^{(i+1)}) - f(\boldsymbol{\theta}^{(i)})| / f(\boldsymbol{\theta}^{(i)}) < \varepsilon$

$$\boldsymbol{\theta}_{opt} = \boldsymbol{\theta}^{(i+1)}, \text{ Terminate the algorithm}$$

else $i \leftarrow i+1$, goto Step 2

Fig. 2. The proposed algorithm

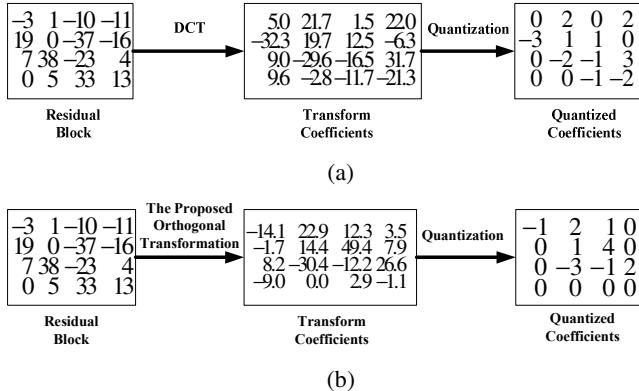


Fig. 3. Comparsion of DCT and the proposed transformation

Fig. 3. gives an example to demonstrate the proposed idea and algorithm against the DCT. Given the same block from MC-residual, after DCT and quantization, we need to transmit 11 non-zero coefficients (Figure 3(a)). The proposed transform results in a more sparse coefficient matrix as in Figure 3(b), and then after the same quantization procedure we have only 8 non-zero coefficients to transmit. We also calculate the Mean Squared Error (MSE) after quantization, the MSE is 31.4 for Figure 3(a) while that is only 26.4 for Figure 3(b). Therefore, we can expect that the transform derived by the proposed algorithm will have higher coding efficiency than DCT and the overall experiment results in Section 4 will confirm this.

Figure 4 shows the convergence performance with the proposed L1-Norm minimization algorithm. We pick the second MC-residual image of CIF video “Mobile” for demonstration. This MC-residual is derived with 8x8 Quarter-pixel-resolution full-search motion-estimation. After 8x8 block based 2D-DCT, the

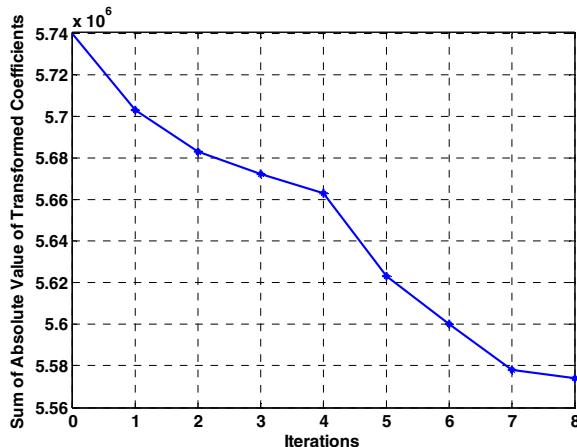


Fig. 4. Illustration of convergence performance with the proposed algorithm

summation of absolute values of DCT coefficients is 5.7488×10^5 . After 8 iterations of L1-Norm minimization, the algorithm converges at the local minimum with 5.5776×10^5 . Although the improvement is only 3% here, but it already accounts for 10% of bitrate saving in Bjontegaard-Delta (BD)-Bitrate metric [10] as we can see in the next section.

4 Overall Experiments and Results

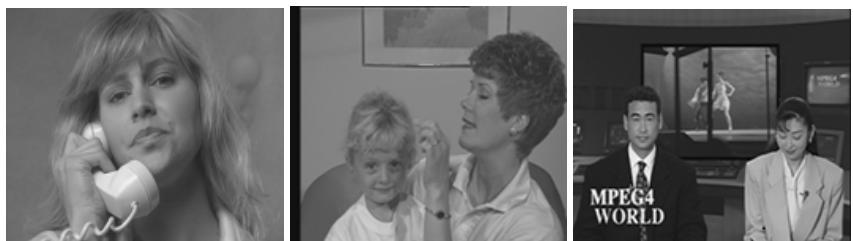
In order to show the effectiveness of the proposed L1-Norm minimization based MC-residual coding, we implement the proposed scheme in H.264/AVC reference software [11] by the substitution of DCT. To provide a comprehensive comparison, 12 test sequences are selected in our experiments, as shown in Fig. 5. We use 6 CIF sequences and 6 QCIF sequences with 180 frames for each at 30 Frame-Per-Second (FPS). These sequences contain various camera and object motion patterns. Five of these sequences contain higher motion activities, e.g. *Foreman* (containing active close-up object and panning background), *Coastguard* (containing moving object and moving background with camera shift), *Football* (containing very fast motion), *Mobile and Bus* (containing moving objects and complex background); other test sequences contain slow motion, e.g. *Akiyo* and *Silent* (containing slow foreground movements), *Suzie* and *Mother* (containing fast foreground movements), *Waterfall* (containing camera zooming), *News* (containing foreground head & shoulder movements and moving object in the background), and *Flower* (containing low motion camera shift but complex objects).

Here we list some of the important encoder parameters. The Group of Pictures (GOP) is set as 30. The first frame is encoded as an I-frame, and all the remaining frames are coded as P-frames. Quarter-pixel-resolution full-search motion-estimation with 8x8 blocks is used. Entropy coding is performed with context-adaptive variable-length-codes (CAVLC). The optimal rotation vector θ_{Opt} is quantized into 8 bits for each rotation as the side information. In our case, for 8x8 orthogonal basis, we need to transmit 28 rotations as the side information for each MC-residual image. With CIF resolution, this side information will only account for 0.002 bpp, which is almost negligible compared with the side information used in [4],[5],[6].

Rate-Distortion (R-D) curves are plotted in Fig. 6 for four video sequences, while the overall R-D performance is summarized in Table 1 for all video sequences, with BD-PSNR and BD-Bitrate for fair comparison cross the bitrates [10]. Table 1 indicates the average BD-Bitrate savings (in percentage) and the average BD-PSNR improvement of the proposed method against and the conventional 2-D DCT one (i.e., H.264/AVC). It is worth noting that the positive number for BD-PSNR means the BD-PSNR improvement, while the negative number for BD-Bitrate means the BD-Bitrate saving, for the proposed scheme.

Ideally, we should also design a new entropy coding table since the distribution of transform coefficients changes after different transformations. However, even we embedded the proposed scheme inside the H.264/AVC codec which is optimized for DCT based coding, our method can still outperform for each test sequence, and this is significant because the test sequences used here are with diversifying visual content and motion. This also represents room for further improvement with the proposed method when entropy coding is customized with the proposed transform.

QCIF Sequences

(a) *Akiyo*(b) *Coastguard*(c) *Mobile*(d) *Suzie*(e) *Mother*(f) *News*

CIF Sequences

(g) *Bus*(h) *Waterfall*(i) *Flower*(j) *Foreman*(k) *Football*(l) *Silent***Fig. 5.** Testing Sequences

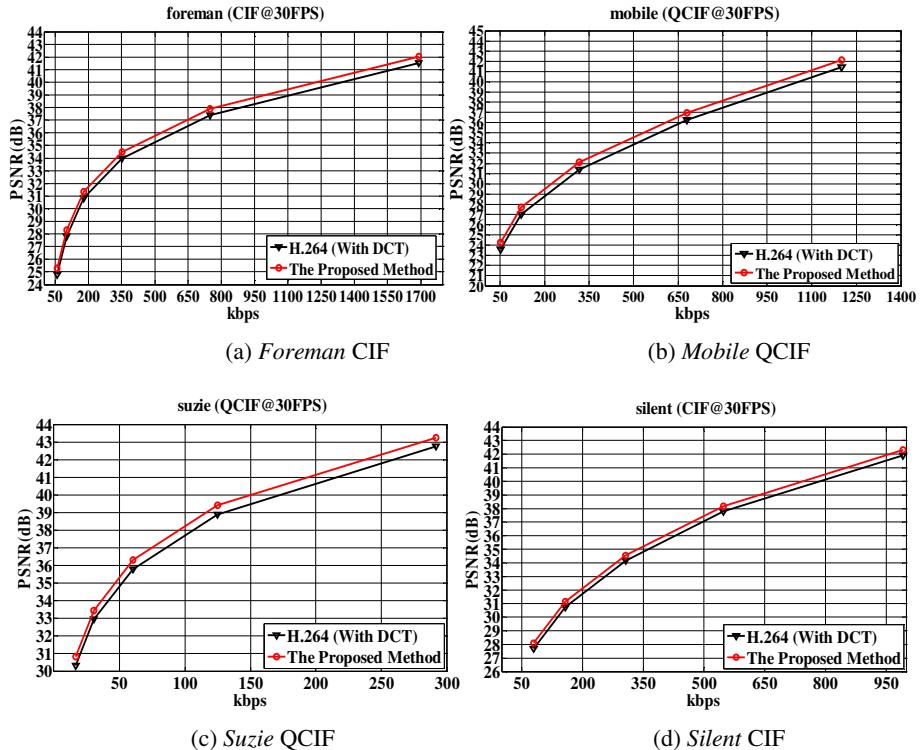


Fig. 6. Several R-D Performance Curves

Table 1. BD-PSNR and BD-Bitrate between the proposed algorithm and the H.264/AVC

Resolution	Sequences	BD-PSNR (dB)	BD-Bitrate (%)
QCIF	Coastguard	0.43	-8.64
	News	0.40	-10.61
	Mobile	0.57	-10.88
	Suzie	0.43	-8.72
	Akiyo	0.23	-3.45
	Mother	0.39	-7.87
CIF	Flower	0.51	-10.40
	Foreman	0.41	-8.72
	Football	0.45	-9.13
	Silent	0.27	-4.03
	Waterfall	0.37	-9.03
	Bus	0.46	-8.78
	<i>Average</i>	0.41	-8.23

As can be seen from the overall performance indicated in Table 1, the BD-PSNR for each sequence increases 0.41dB on average by the proposed algorithm compared to the state-of-the-art technology for video coding (i.e., H.264/AVC), or equivalently the proposed algorithm achieves BD-Bitrate saving up to 10.88% with an average of 8.23%.

In the proposed method, there is only slight increase (5-8%) in the computational time due to the iteration algorithm. During each iteration, we only need to evaluate the performance of the orthogonal matrix without quantization. According to the experiment in [12], for H.264/AVC encoder, the operations for DCT + Quantization (Q) + Inverse-Quantization(IQ) + Inverse-DCT(IDCT) + Motion Compensation(MC) only occupy 0.45% of the overall computational time. Since our iteration algorithm only repeats for the transformation step, which is similar to the DCT step in the conventional encoder in terms of complexity, the computational overhead introduced by the iteration algorithm is minor compared with other component (e.g. Motion Estimation (ME) and interplotaion take more than 95% of overall computational time [12]).

It is also worth noting that the proposed method gives higher coding efficiency improvement for video sequences with high/complex motion (e.g. *Mobile* 0.57dB, *Flower* 0.51dB, *Bus* 0.46dB). This is because for low/simple motion videos, most of the MC-residual pixels are zeros, for which DCT is good enough to compress and leaves little room for our method to improve; while for high/complex motion videos, there are many more non-zeros MC-residual pixels and this gives our method more room to improve the coding efficiency. This is a very desirable attribute for video compression. When the motion becomes high/complex, the bitstream size after coding often increases dramatically, while our method provides more bitrate saving (in percentage) for complex motion. Both of these facts imply that our method has the potential to save more bits for high/complex motion videos. The proposed method and its advantage are particularly significant since it is more challenging to deal with high/complex motion for the existing video coding schemes; a new scheme capable of outperforming the state-of-the-art codec with high/complex motion is much more meaningful than an one that is better with low/simple motion, because the former situation of motion is the bottleneck of the reduction in bitrate and its fluctuation.

5 Conclusion

In video coding, motion compensation (MC) residual has different characteristics from a natural image, and this makes the current compression schemes less efficient with complex motion. Therefore, it is desirable to develop a transform that is adapted to the residual to yield coefficient compaction, while keeping the necessary side information minimum. In this paper, we propose a L1-Norm minimization based algorithm to generate an orthogonal basis for coding the MC-residual image. By converting the orthogonal constraints into convex constraints, we are able to conduct the gradient-based search. By setting the DCT matrix as the starting searching matrix, we are guaranteed to obtain a better result in terms of L1-Norm minimization; in addition, this combined with the use of rotation matrices avoids transmission of substantial side information (necessary for the adaptive transform matrices). Experiment results indicate that using the orthogonal transformation derived from the proposed algorithm, we can improve the coding efficiency of the MC-Residual for

video with diversified visual content and motion. The better outperformance of the proposed transform at high/complex motion is a highly desirable characteristic of video coding.

References

1. Niehsen, W., Brunig, M.: Covariance analysis of motion-compensated frame differences. *IEEE T-CSVT* 9(4), 536–539 (1999)
2. Chen, F., Pang, K.: The optimal transform of motion-compensated frame difference images in a hybrid coder. *IEEE T-CAS2* 40(6), 393–397 (1993)
3. Hui, K.-C., Siu, W.-C.: Extended analysis of motion-compensated frame difference for block-based motion prediction error. *IEEE T-IP* 16(5), 1232–1245 (2007)
4. Waldemar, P., Ramstad, T.A.: Hybrid KLT-SVD Image Compression. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, vol. 4, pp. 2713–2716 (1997)
5. Dapena, A., Ahalt, S.: A Hybrid DCT-SVD Image-Coding Algorithm. *IEEE T-CSVT* 12, 114–121 (2002)
6. Ochoa, H., Rao, K.R.: A Hybrid DWT-SVD Image-Coding System (HDWTSVD) for Monochromatic Images. *WSEAS Transaction on Circuit and Systems* 4, 419–424 (2005)
7. Donoho, D.L., Elad, M.: Maximal sparsity representation via ℓ_1 minimization. *Proceedings of the National Academy of Sciences* 100(5), 2197–2202 (2003)
8. Kwak, N.: Principal component analysis based on L_1 -norm maximization. *IEEE T-PAMI* 3(9), 1672–1680 (2008)
9. Pang, Y., Li, X., Yuan, Y.: Robust Tensor Analysis with L_1 -Norm. *IEEE T-CSVT* 20(2), 172–178 (2010)
10. Bjontegaard, G.: Calculation of average PSNR differences between rd-curves, VCEG Contribution VCEG-M33
11. H.264/AVC Reference Software, JM14.0:
<http://iphome.hhi.de/suehring/tm1/>
12. Huang, Y., Hsieh, B., Chien, S., Ma, S., Chen, L.: Analysis and complexity reduction of multiple reference frames motion estimation in h.264/AVC. *IEEE T-CSVT* 16(4), 507–522 (2006)

Parallel Deblocking Filter for H.264/AVC on the TILER A Many-Core Systems

Chenggang Yan^{1,2}, Feng Dai¹, and Yongdong Zhang¹

¹ Multimedia Computing Group, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, China

² Graduate University of Chinese Academy of Sciences, Beijing
`{yanchenggang, fdai, zhyd}@ict.ac.cn`

Abstract. For the purpose of accelerating deblocking filter, which accounts for a significant percentage of H.264/AVC decoding time, some studies use waveform method to achieve the required performance on multi-core platforms. We study the problem under the context of many-core systems and present a new method to exploit the implicit parallelism. We apply our implementation to the deblocking filter of the H.264/AVC reference software JM15.1 on a 64-core TILER A and achieve more than eleven times speedup for 1280*720(HD) videos. Meanwhile the proposed method achieves an overall decoding speedup of 140% for the HD videos. Compared to the waveform method, we also have a significant speedup 200% for 720*576(SD) videos.

Keywords: Video decoding; H.264/AVC; Deblocking filter; Parallel algorithm; Many-core systems.

1 Introduction

H.264/AVC is the newest coding standard which brings a high efficient video compression method to the multimedia industry [1]. Due to its compression efficiency, it has been applied to several important applications including video telephony, video storage, broadcast, video streaming, HD-DVD. An in-loop deblocking filter has been adopted by H.264/AVC, though provides 10–15% bit rate saving, brings heavy computation [2]. For high definition videos, Chen et al. [3] shows that the deblocking filter contributes 38% computation time. Therefore, it is important to improve the H.264/AVC deblocking performance.

With the development of IC technologies, multi-core processors are commonplace and many researchers have tried to use multi-core platforms to implement parallelization for H.264/AVC decoder [4–6]. The waveform method is a commonly used MB-level data partition [6,7]. As the number of cores per chip increases, we are entering into many-core ages such as TILER A many-core systems. It is not easy to develop an efficient data partition schedule for deblocking filter that can efficiently utilize so many cores. In order to solve such problem, we review the deblocking filter within MBs. We study the data dependencies in low level and change the edge filtering orders. It leads to a significant change of data dependencies between neighboring MBs, while the final results are not changed. Based on this modification we can divide the deblocking filter into more tasks executed on many-core processors.

The remainder of this paper is organized as follows. First, we briefly introduce the TILERa Many-core systems in Section 2. An overview of H.264/AVC deblocking filter and the wavefront method are described in Section 3. The proposed deblocking filter speedup techniques are presented in Section 4. Experimental results and analysis are given in Section 5. Finally, we conclude this letter in Section 6.

2 Short Overview of TILERa Many-Core Systems

Centralized monolithic processor designs, such as single core and shared bus structures, are not scaling, leading to multi-core design becoming the norm [8-10]. Research highlights the benefits of mesh networks connecting these cores [11,12], which can solve the multi-core scalability problem.

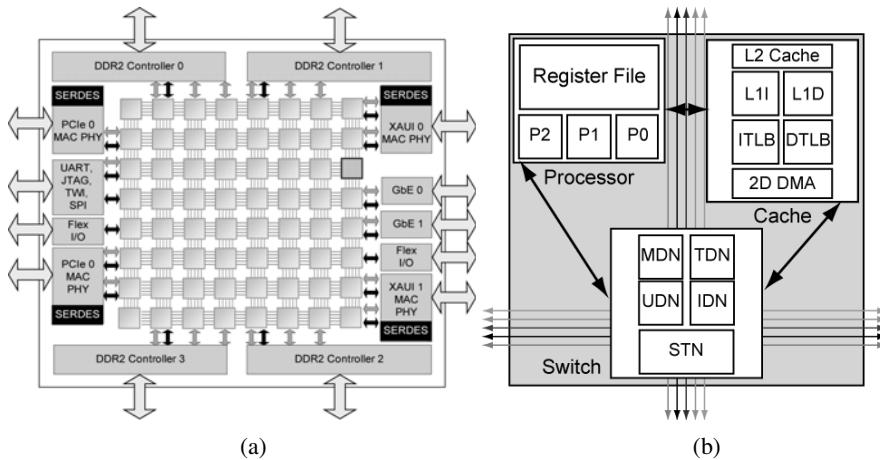


Fig. 1. 64-core TILERa block diagram and Single tile block diagram figure (a) 64-core TILERa (b) single tile

TILERa many-core processor family features devices with 16 to 100 identical processor cores. Fig.1a illustrates a block diagram with 64-core TILERa arranged in an 8x8 array. These cores run at 700MHz and connect through a scalable and high-speed 2D mesh network. Using four 72-bit 800MHz DDR2 interfaces, the chip peak memory bandwidth is over 25GB/s. Two PCI-e $\times 4$ interfaces, two XAUI interfaces and two RGMII interfaces provide over 40Gb/s of I/O bandwidth. The low-speed interfaces provide seamless integration with a variety of systems [13].

Each tile processor (Fig.1b) is a full featured processor and provides a three-way Very Long Instruction Word (VLIW) architecture allowing up to 3 instructions per cycle. A single core contains a 16KB L1 cache, 8KB of data L1 cache, and 16KB of combined L2 cache. Each core also has its own DMA engine and Translation Lookaside Buffer (TLB) which allow memory virtualization and the ability to run a modern O/S on each core [14].

3 Deblocking Filter and the Wavefront Method

H.264/AVC video coding standard adopts an in-loop deblocking filter which removes block-edge artifacts that are produced by block transformation and block motion compensation. “In-loop” implies any inappropriate modification of the in-loop deblocking filter causes serious error propagation to succeeding frames.

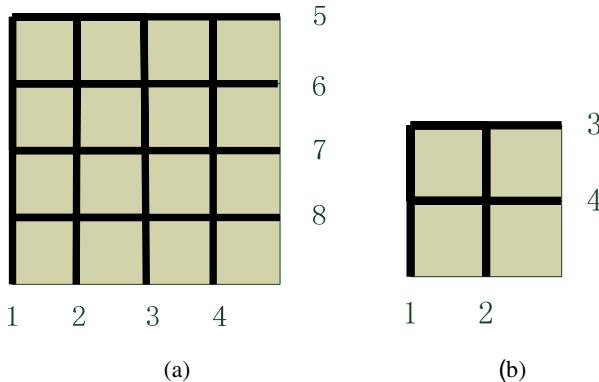


Fig. 2. Edge filtering order in each MB (a) 16x16 Luminance data (b) 8x8 Chrominance data

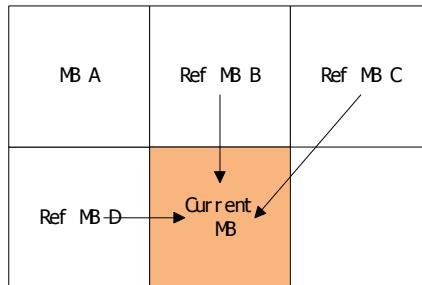
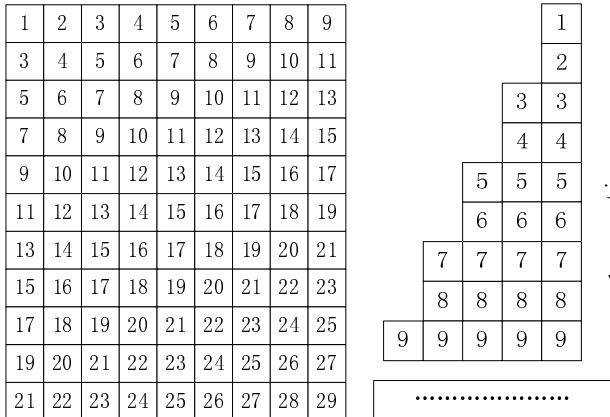


Fig. 3. Data dependencies between neighboring MBs in deblocking filter

The deblocking filter stage is applied to each luminance and chrominance edge within one MB (Fig. 2). All edges are processed from left to right and from top to bottom. Vertical boundary edges are processed first. Usually MBs in a frame are processed in scan order during deblocking filter which should be followed to ensure both the encoder and the decoder having the same result. The current deblocking MB has dependencies on its adjacent left, upper, and upper-right MBs (Fig. 3). When deblocking the current MB, the left, upper, and upper-right MB should have completed processed.

1	2	3	4	5
3	4	5	6	7
5	6	7	8	9

Fig. 4. A wavefront data partition example: each rectangular indicts an MB**Fig. 5.** An example of wavefront MB-Level deblocking filter for a video frame

The wavefront method is widely used for data partition [7] which processes the data in wavefront order(Fig.4). The MBs are processed according to their numbers which indicate the time stamp. MBs with the same numbers are processed concurrently. Therefore, some researchers adopt this method to parallel deblocking filter. In this paper we call it wavefront MB-level deblocking filter method and wavefront method for short. Fig.5 shows an example of wavefront method for a video frame. At time stamp 4, there are two independent MBs. The maximum number of independent MBs(MNIM) is first available at time stamp 9 and it depends on the resolution. Table1 states the MNIM using the wavefront method. The column “MBs” indicates the horizontal and vertical numbers of MBs in a frame.

There are some disadvantages for this kind of MB-level parallelism. The first disadvantage is that there are few numbers of independent MBs at the start of deblocking filter. Fig.5 shows that the MNIM increases one at two stamps addition. What's more, when the number of cores is enough, this method has not discovered the potential capacity of the MNIM. As shown before the MNIM increases with the resolution of the frame. Considering the number of cores, we can calculate the maximum parallelism as follows:

$$MP_{wavefront} = \begin{cases} \min(\text{ceil}(\frac{W}{2}), H) & \text{if } \min(\text{ceil}(\frac{W}{2}), H) < C \\ C & \text{if } \min(\text{ceil}(\frac{W}{2}), H) \geq C \end{cases} \quad (1)$$

Where ceil function returns the value of a number rounded upwards to the nearest integer, \min function returns smaller value of the two integers, H and W indicates the horizontal and vertical numbers of MB in a frame, C indicates the number of cores used for deblocking filter.

Table 1. The MNIM using the wavefront method

Format	Resolution	MBs	MNIM
QCIF	176x144	11x9	5
CIF	352x288	22x18	9
SD	720x576	45x36	18
HD	1280x720	80x45	23
FHD	1920x1088	120x68	34

Another drawback of the wavefront method is that we should not ignore the communication between the MBs. When the current MB is processed, it has to communicate with up to three MBs processed on other cores. The amount communication overhead incurred in wavefront method is approximately $3*W*H$. Assuming the deblocking time of each MB is identical and the number of cores is enough. We can get the speedup of data parallelism in wavefront method as follows:

$$SP_{wavefront} = \frac{H * W}{H + 2 * (W - 1) + 3 * k * W * H} \quad (2)$$

Where k indicates the time it takes to communicate between two MBs.

4 The Proposed MB-Level Deblocking Filter

As described before, the order of deblocking filter should not change arbitrarily which can influence the results of H.264/AVC decoder. MBs can be processed out of scan order provided these dependencies are satisfied. The wavefront method studies the dependencies in MB-Level. In this section we study it in low level.

Fig. 6 shows the data dependencies between the current luminance MB and edges in neighboring MBs. As shown in Fig. 6, the four left neighbour 4x4 blocks of current MB are named L1-L4; the four upper neighbour 4x4 blocks of current MB are named T1-T4; V1-V7 are the vertical edges; H1-H6 are the horizontal edges; V1, V5, V7, H1 and H5 are edges between the MBs. Current MB can be processed when the neighbor blocks L1-L4 and T1-T4 don't change.

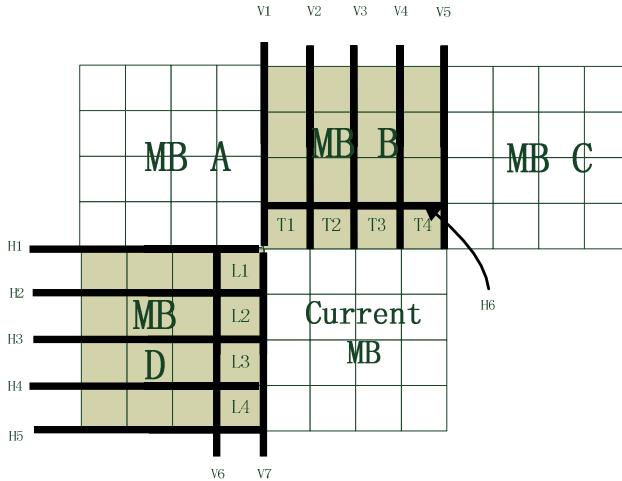


Fig. 6. Data dependencies between the current Luminance MB and neighboring edge filters in deblocking filter

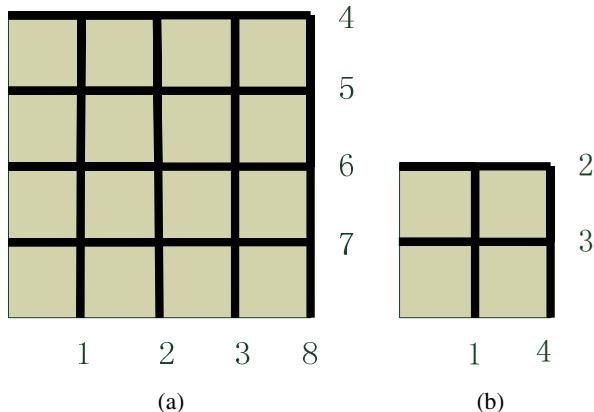


Fig. 7. Edge filtering order in our proposed method (a) 16x16 Luminance data (b) 8x8 Chrominance data

For edge filters, at most four pixels on either side of each edge are evaluated. Up to three pixels on either side of the edge may be influenced[1]. So when the V1-V5 and H6 have the conduction of deblocking filter, the T1-T4 will not change; when the H1-H4 and V6 are processed, the L1-L4 will not alter, as well. Though H5 could change the result of L4, it happens after the deblocking filter of current MB based on the global deblocking order for MBs. Therefore, it will not influence the current MB. MB C has impact on the current MB only through the vertical edge V5. If the vertical edge V5 belong to the deblocking filter of MB B, the MB C is irrelevant to current MB. Meanwhile V1 belongs to MB A and V7 belongs to MB D by this rule. In this

condition, it seems that the current MB has an additional relevance MB A. But when the MB B and MB C are processed, the MB A must have been ready. So the current MB has nothing to do with the MB A.

Based on this observation we proposed a new MB-Level deblocking filter. First of all, we changes the edge filtering order within an MB(Fig. 7) but not the overall edge filtering order in one frame. The data dependencies between neighboring MBs in deblocking filter decreases (Fig. 8). The current deblocking MB only has dependencies on its adjacent left and upper MBs. Assuming every MB adopts the edge filtering order we proposed, we can have a new method of data partition. Fig. 9 shows an example of the proposed data partition for a video frame. At time stamp 4, there are four independent MBs. The MNIM depends on the resolution as the wavefront method. Table2 states the MNIM using our proposed MB-level deblocking filter.

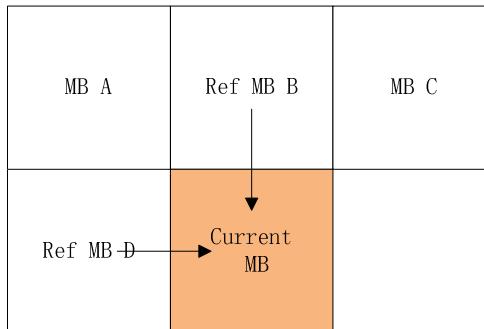


Fig. 8. Data dependencies between neighboring MBs in our proposed method

Table 2. Maximum number of independent MBs using our proposed MB-level deblocking

Format	Resolution	MBs	MNIM
QCIF	176x144	11x9	9
CIF	352x288	22x18	18
SD	720x576	45x36	36
HD	1280x720	80x45	45
FHD	1920x1088	120x68	68

Compared with the wavefront method, there are some advantages for this kind of MB-level parallelism. The first advantage is that there are more numbers of independent MBs at the start of deblocking filter. Fig. 9 shows that the MNIM increases one at one stamp addition. What's more, when the number of cores is enough, the MNIM is bigger (Table 2). Considering the number of cores, we can calculate the maximum parallelism as follows:

$$MP_{new} = \begin{cases} \min(W, H) & \text{if } \min(W, H) < C \\ C & \text{if } \min(W, H) \geq C \end{cases} \quad (3)$$

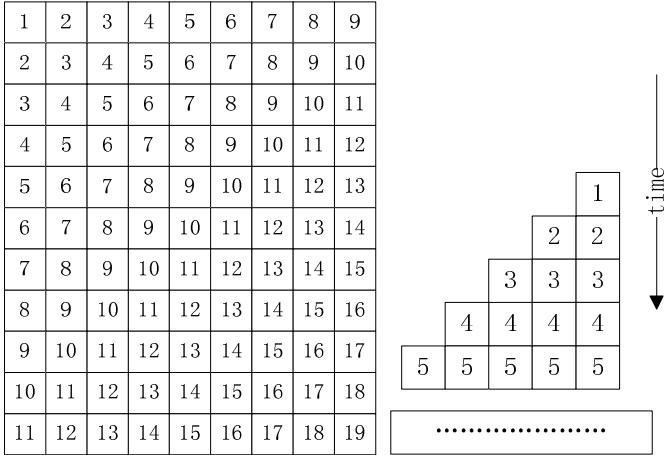


Fig. 9. An example of our proposed MB-Level deblocking filter for a video frame

Obviously equation (1) and (3) shows that with enough cores, the maximum parallelism of our method is better. Further each MB in our method to be processed has at most two connections with others. The amount of communication overhead incurred in our proposed MB-level deblocking filter is $2*W*H$ approximately which is less than the wavefront method. When the number of cores is enough, we can also get the speedup of data parallelism in our proposed method as follows:

$$SP_{proposed} = \frac{H * W}{H + W - 1 + 2 * k * W * H} \quad (4)$$

Where k indicates the time it takes to communicate between two MBs.

From equation (2) and (4), we find that the theoretical speedup in our method is better than that in wavefront method. In the next section, we will compare them from actual experiments.

5 Experimental Results

In this section, we compare our scheme on deblocking filter with the wavefront method. To compare the two methods, we adopted a decoder migrated from H.264/AVC reference software JM15.1 without any optimization as baseline profile. The decoder includes the stages of entropy decoding, de-quantization, inverse integer transform, intra prediction and motion compensation. We implemented the parallel algorithm on a 64-core TILER A which was described in section 2.

The input videos in our experiments contain a list of standard test sequences named *mobile*, *highway*, *hall*, *mother-daughter*, *riverbed*, *rush_hour* *pedestrian*, *container* and *blue_sky* with four resolution levels, 1280x720(HD), 720x576(SD), 352x288(CIF) and 176x144(QCIF), which are encoded by the reference software JM15.1. The speedup of deblocking filter which is gained by our proposed method compared to the wavefront method can be measured by:

$$Speedup = \frac{Wavefront(ms)}{Proposed(ms)} \quad (5)$$

Table 3. The execution time between wavefront MB-level deblocking filter and our proposed method

sequences	Format	JM15.1(ms)	Wavefront method(ms)	Proposed method(ms)	Speedup
blue_sky	SD	268.8	49.7	23.8	2.09
blue_sky	HD	596.5	76.5	42.7	1.79
pedestrian	SD	225.1	45.3	22.9	1.98
pedestrian	HD	526.7	75.3	43.0	1.75
riverbed	SD	168.3	35.4	19.4	1.82
riverbed	HD	382.9	69.2	40.6	1.70
rush_hour	SD	212.4	46.7	22.7	2.06
rush_hour	HD	508.5	76.7	43.9	1.75
container	QCIF	15.8	8.6	4.2	2.05
container	CIF	55.5	18.6	9.3	2.00
hall	QCIF	15.3	8.4	4.2	2.00
hall	CIF	61.7	19.5	9.6	2.03
highway	QCIF	13.4	7.5	3.8	1.97
highway	CIF	58.9	18.6	9.2	2.02
mobile	QCIF	11.1	6.6	3.2	2.06
mobile	CIF	46.0	16.7	8.2	2.04
mother-daughter	QCIF	14.4	6.9	3.9	1.77
mother-daughter	CIF	60.6	19.4	9.6	2.02

Table 4. The average execution time between wavefront MB-level deblocking filter and our proposed method

Format	JM15.1(ms)	Wavefront method(ms)	Proposed method(ms)	Speedup
QCIF	14.0	7.6	6.3	1.21
CIF	56.5	18.6	12.4	1.50
SD	218.7	44.3	22.2	2.00
HD	503.7	74.4	42.6	1.75

Table 3 compares the performance of deblocking filter between wavefront method and our proposed method. The average execution time is summarized in Table 4. Results demonstrate that the proposed scheme can achieve more than eleven times speedup for HD videos compared to the JM15.1 software. We also can have a speedup of 200% for SD videos compared to the wavefront method.

Table 5 shows the overall decoding speedup compared with the JM15.1 software. From Table 6, we find that the proposed method achieves better speedup as the resolution of videos increases. The proposed method achieves a speedup of 140% for the HD videos.

Table 5. The overall decoding speedup

sequences	Format	JM15.1(ms)	Proposed method(ms)	Speedup
blue_sky	SD	777	560	1.39
blue_sky	HD	1621	1131	1.43
pedestrian	SD	651	477	1.36
pedestrian	HD	1396	949	1.47
riverbed	SD	687	542	1.27
riverbed	HD	1449	1143	1.27
rush_hour	SD	629	468	1.34
rush_hour	HD	1369	941	1.45
container	QCIF	69	60	1.15
container	CIF	203	170	1.19
hall	QCIF	67	59	1.14
hall	CIF	201	163	1.23
highway	QCIF	58	51	1.14
highway	CIF	177	141	1.26
mobile	QCIF	85	79	1.08
mobile	CIF	258	224	1.15
mother-daughter	QCIF	63	55	1.15
mother-daughter	CIF	184	146	1.26

Table 6. The average overall decoding speedup

Format	JM15.1(ms)	Proposed method(ms)	Speedup
QCIF	68.4	60.8	1.13
CIF	204.6	168.8	1.21
SD	686.0	511.8	1.34
HD	1458.8	1041.0	1.40

6 Conclusion

The deblocking filter is considered as the most computationally expensive part of the H.264/AVC decoder. A new MB-level parallelism is proposed to speedup deblocking filter which is better than the widely used MB-level data partition schedule named wavefront method. The proposed scheme was shown to reduce the deblocking computational load by more than eleven times for HD videos in comparison with the reference software JM15.1. It also has a speedup of 200% for SD videos compared to the wavefront method. The overall decoding time is reduced significantly, as well.

Acknowledgments

This work was supported by the National Nature Science Foundation of China (60802028, 60873165), National Basic Research Program of China (973Program, 2007CB311100), Co-building Program of Beijing Municipal Education Commission, Beijing New Star Project on Science & Technology (2007B071).

References

1. Joint Video Team of ITU-T and ISO/IEC JTC1. Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification. Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVTC050 (2003)
2. List, P., Joch, A., Lainema, J., Bjntegaard, G., Karczewicz, M.: Adaptive deblocking filter. *IEEE Transactions on Circuits and Systems for Video Technology* 13(7), 614–619 (2003)
3. Chen, T.C., Fang, H.C., Lian, C.J., Tsai, C.H., Huang, Y.W., Chen, T.W., et al.: Algorithm analysis and architecture design for HDTV applications-a look at the H. 264/AVC video compressor system. *IEEE Transactions on Circuits and Devices Magazine* 22(3), 22–31 (2003)
4. Zhao, Z., Liang, P.: Data partition for waveform parallelization of H.264 video encoder. In: *IEEE International Symposium on Circuits and Systems, ISCAS 2006*, pp. 21–24 (2006)
5. Lee, J.-Y., Lee, J.-J., Park, S.M.: Multi-core platform for an efficient H.264 and VC-1 video decoding based on macroblock row-level parallelism. *IET Circuits, Devices & Systems* (2010)
6. Meenderinck, C., Azevedo, A., Alvarez, M., Juurlink, B., Mesa, M.A., Ramirez, A.: Parallel Scalability of Video Decoders. Delft University of Technology (2008)
7. Aho, A.V., Sethi, R., Ullman, J.D.: *Compilers: principles, techniques, and tools*. Addison-Wesley Longman, Boston (2007)
8. Friedrich, J., McCredie, B., James, N., et al.: Design of the Power6™ Microprocessor. *ISSCC Dig. Tech. Papers*, pp. 96–97 (2007)
9. Dorsey, J., Searles, S., Ciraula, M., et al.: An Integrated Quad-Core™ Opteron Processor. *ISSCC Dig. Tech. Papers*, pp. 102–103 (2007)
10. Nawathe, U., Hassan, M., Warriner, L., et al.: An 8-Core 64-Thread 65b Power-Efficient SPARC SoC. *ISSCC Dig. Tech. Papers*, pp. 108–109 (2007)
11. Taylor, M., Kim, J., Miller, J., et al.: A 16-Issue Multiple-Program-Counter Microprocessor with Point-to-Point Scalar Operand Network. *ISSCC Dig. Tech. Papers*, pp. 170–171 (2003)
12. Vangal, S., et al.: An 80-tile 1.28TFLOPS Network-on-Chip in 65nm CMOS. *ISSCC Dig. Tech. Papers*, p. 98 (2007)
13. Agarwal, A., Bao, L., Brown, J., et al.: Tile Processor: Embedded Multicore for Networking and Digital Multimedia. *Hot Chips* (2007)
14. Bell, S., Edwards, B., Amann, J., et al.: TILE64-Processor: A 64-Core SoC with Mesh. In: *Interconnect Solid-State Circuits Conference* (2008)

Image Distortion Estimation by Hash Comparison

Li Weng and Bart Preneel*

Katholieke Universiteit Leuven, ESAT/COSIC-IBBT

li.weng@esat.kuleuven.be, bart.preneel@esat.kuleuven.be

Abstract. Perceptual hashing is conventionally used for content identification and authentication. In this work, we explore a new application of image hashing techniques. By comparing the hash values of original images and their compressed versions, we are able to estimate the distortion level. A particular image hash algorithm is proposed for this application. The distortion level is measured by the signal to noise ratio (SNR). It is estimated from the bit error rate (BER) of hash values. The estimation performance is evaluated by experiments. The JPEG, JPEG2000 compression, and additive white Gaussian noise are considered. We show that a theoretical model does not work well in practice. In order to improve estimation accuracy, we introduce a correction term in the theoretical model. We find that the correction term is highly correlated to the BER and the uncorrected SNR. Therefore it can be predicted using a linear model. A new estimation procedure is defined accordingly. New experiment results are much improved.

1 Introduction

Perceptual hash (PH) algorithms are a particular category of hash functions for the multimedia domain [1,2,3]. They generate hash values based on robust features of multimedia data. A hash value is typically a compact binary string. Differing from cryptographic hash values[4,5], a PH value is only based on the content. It is usually insensitive to incidental distortion, such as compression, scaling, and photometric processing. Therefore it can be used as a persistent fingerprint of the corresponding content. One can efficiently find similar multimedia content by comparing PH values. Therefore, PH algorithms are useful tools for applications such as multimedia databases and multimedia search engines.

* This work was supported in part by the Concerted Research Action (GOA) AMBioRICS 2005/11 of the Flemish Government and by the IAP Programme P6/26 BCRYPT of the Belgian State (Belgian Science Policy). The first author was supported by the IBBT/AQUA project. IBBT (Interdisciplinary Institute for BroadBand Technology) is a research institute founded in 2004 by the Flemish Government, and the involved companies and institutions (Philips, IPGlobalnet, Vitalsys, Landsbond onafhankelijke ziekenfondsen, UZ-Gent). Additional support was provided by the FWO (Fonds Wetenschappelijk Onderzoek) within the project G.0206.08 Perceptual Hashing and Semi-fragile Watermarking.

The usage of perceptual hashing has been evolving. It is usually used for multimedia content identification [1]. It is also used for generating content-based watermarks [2]. New PH algorithms are sometimes designed to support a secret key [3]. The hash value being highly dependent on the key enables message authentication. Compared with conventional message authentication codes (MAC) or digital signatures (DS) [4,5], *perceptual* MAC and DS do not need to be regenerated when the multimedia data undergoes incidental distortion. Some PH algorithms also support tamper localization [6,7,8]. It seems that PH algorithms are having growing potential for new applications.

Recently, a new application of perceptual hashing has been found. Besides content identification and authentication, perceptual hashing can also help with quality assessment or distortion estimation. Since the hash value is related to the content, quality degradation might reflect from the hash value. Given a multimedia file and its distorted (compressed) version, one can tell the distortion level by comparing their PH values. This idea was successfully demonstrated by Doets and Lagendijk for estimating music compression distortion [9]. An interesting question is whether the same idea works for images or video. In this work, we use an image hash algorithm for image distortion estimation. We show that it is possible to estimate the signal to noise ratio according to the hash distance. The considered distortions are the JPEG compression, the JPEG2000 compression, and the additive white Gaussian noise (AWGN). The rest of the work is organized as follows: Section 2 introduces the background of image distortion estimation; Section 3 proposes an image hash algorithm, and introduces a model for distortion estimation; Section 4 shows experiment results; Section 5 proposes a method to improve estimation accuracy; Section 6 concludes the work.

2 Image Distortion Estimation

Image distortion estimation is a conventional research topic. It is closely related to image quality assessment and image tamper detection/localization. There have been many approaches which generally divide into two categories – ones that require the original data, e.g., [10,11], and ones that do not. The former is generally more accurate, while the latter is more practical (since original data is not always available). Recently, it was found that hash comparison can also indicate the distortion to some extent [9] – a new feature offered by perceptual hashing. This approach is essentially a trade-off between the above two categories. Since a hash value is related to the content, distortion also influences the hash value. A typical user case is the distribution chain of high quality multimedia content. In a heterogeneous network, the quality of content generally degrades from the source/server (e.g. a production studio) to the end-user/client (e.g. a mobile device). As a content provider, it is often necessary to monitor the quality (degradation) along the distribution chain. Although parameters such as the bit rate can tell the quality to some extent, they are generally not reliable. It is better to assess the quality on the client side and report to the server. Due to many reasons, e.g., bandwidth or copyright limit, original multimedia data

cannot be available to the end-users. Instead, sending hash values is a good alternative. End-users could compute hash values from the received content and compare them with the original ones to estimate the quality. Figure 1 illustrates this idea. Besides distortion estimation, the availability of hash values also makes it easy to perform content-based search or content authentication.

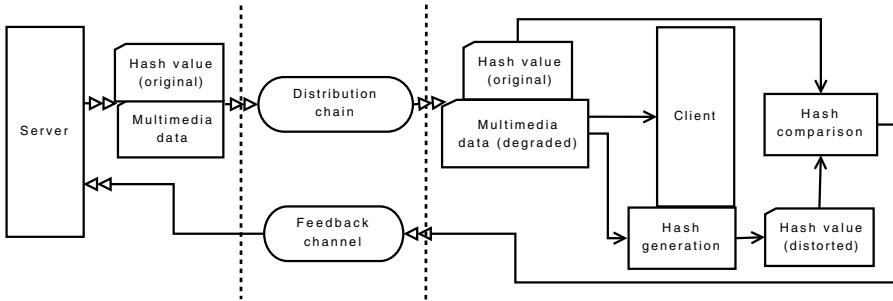


Fig. 1. An example application scenario of distortion estimation

More specifically, distortion estimation by hash comparison also has two kinds of approaches. Some only use the hash values; some also make use of the received multimedia data. Intuitively, utilizing both the hash value and the distorted version gives the best results, because more information is available. Tagliasacchi *et al.* have already demonstrated this approach [7]. They consider an optimization problem – given the original hash and the distorted signal, search for a sparse error pattern which results in the distorted hash. However, it is computationally expensive and requires very sophisticated algorithms. In this work, we only exploit hash values. This method is simple and fast. We focus on the estimation of compression distortion. The signal to noise ratio (SNR) is used as the distortion metric. In the following, we first propose an image hash algorithm.

3 An Image Hash Algorithm for Distortion Estimation

We have been inspired by the work of Doets and Lagendijk [9], where they estimate music compression distortion by comparing audio hash values. They used the audio hash algorithm proposed by Haitsma and Kalker [12]. Although it was designed for audio, we find that the basic principles might also apply to images. It is interesting to see how well it works in a different domain. Therefore, we adapted the original audio hash algorithm for images.

The new algorithm is particularly designed for high definition (HD) images, because distortion estimation is more interesting for such content. The algorithm is block-based. A sub-hash value is computed from each block. There are the following steps:

- Pre-processing: the input image is converted to gray-scale and proportionally resized so that the maximum length is 1280 pixels; the mean value is subtracted from the image;

- Block division: the pre-processed image is first padded then divided into 64×64 pixel blocks with 50% overlapping;
- Power spectral density (PSD) estimation: the PSD of block n is estimated from the 2D Fourier transform of the corresponding image block $X_n(k)$:

$$S_n(k) = \frac{1}{L^2} |X_n(k)|^2, \quad L = 64; \quad (1)$$

- PSD extraction: for each block PSD, the low frequency components corresponding to normalized frequencies 0–0.5 are extracted, see Fig. 2;
- Frequency band division: each extracted (2D) PSD is divided into frequency bands of size 16×16 ; the energy of band m is computed by summing up all values in the band:

$$E_{n,m}^b = \sum_{k \text{ in band } m} S_n(k); \quad (2)$$

- Compute difference between frequency bands in each block:

$$ED'_{n,m} = E_{n,m}^b - E_{n,m+1}^b; \quad (3)$$

- Compute difference between blocks:

$$ED''_{n,m} = ED'_{n,m} - ED'_{n+1,m}; \quad (4)$$

- Hash generation:

$$H(n, m) = \begin{cases} 1, & \text{if } ED''_{n,m} > 0 \\ 0, & \text{if } ED''_{n,m} \leq 0 \end{cases}. \quad (5)$$

The final hash value is the concatenation of all sub-hash values.

The canonical size 1280 in the algorithm is relatively large compared with conventional values such as 512. This is because a small canonical size makes the hash value insensitive to slight distortion introduced by compression. In order to suit the particular need of the application, the hash is designed to have slightly less robustness. Since the pre-processing keeps the original aspect ratio, the size of a hash value is not constant. The maximum size is 4563 bits for the case of 1280×1280 pixels. Although it is larger than conventional ones which are below 1000 bits, it is still negligible considering the data volume of HD content.

3.1 SNR Estimation

A model was derived in [9] to describe the relationship between the SNR and the BER. It also turns out to work for the above algorithm. In this section, we briefly introduce this model. Denote the original image signal by x and the distorted one by y . The distortion is modeled as additive noise, denoted by w :

$$y(i) = x(i) + w(i). \quad (6)$$

Following the notation in the previous section, denote the block difference by ED''_X and ED''_Y respectively. In the derivation of the model, the following assumptions are made:

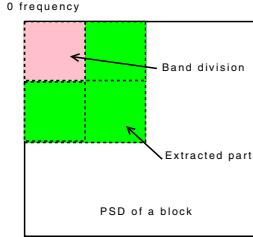


Fig. 2. Frequency band extraction and division

- *Assumption 1:* The signal contribution ED''_X and the noise contribution $ED''_Y - ED''_X$ are zero-mean, mutually independent, normally distributed random variables;
- *Assumption 2:* $x(i)$ and $w(i)$ are zero-mean, normally-distributed random variables with variance σ_X^2 and σ_W^2 .

According to Assumption 1 [9, Appendix 1], we have:

$$P_e = \frac{1}{\pi} \left(\sqrt{\frac{\text{VAR}[ED''_Y - ED''_X]}{\text{VAR}[ED''_X]}} \right) . \quad (7)$$

Assumption 2 gives the following relationship [9, Appendix 2–3]:

$$\text{VAR}[ED''_Y - ED''_X] = \left(\frac{\sigma_W^4}{\sigma_X^4} + 2 \frac{\sigma_W^2}{\sigma_X^2} \right) \text{VAR}[ED''_X] . \quad (8)$$

Defining $\text{SNR} = \frac{\sigma_X^2}{\sigma_W^2}$, we have the relationship between the BER P_e and SNR by combining (7) and (8):

$$P_e = \frac{1}{\pi} \arctan \left(\sqrt{\frac{\sigma_W^4}{\sigma_X^4} + 2 \frac{\sigma_W^2}{\sigma_X^2}} \right) = \frac{1}{\pi} \arctan \left(\sqrt{\frac{1}{\text{SNR}^2} + \frac{2}{\text{SNR}}} \right) . \quad (9)$$

Knowing the BER, solving the above equation is straight-forward:

$$\text{SNR} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, \quad (a = \tan(\pi P_e)^2, b = -2, c = -1) . \quad (10)$$

There is only one meaningful solution.

Equation (10) is a convenient tool to estimate the SNR by BER. Figure 3 shows the ideal relationship. The asterisks in the figure are the BER values computed from synthesized data. In the simulation, we first set a specific SNR, then randomly generate 1280×1280 input images and add noise according to the SNR. Both the input signal and noise are normally distributed with zero mean. For each SNR, we repeat 10 times and plot the average BER value. Clearly, the computed BER values fit the theoretical curve perfectly. Therefore, the proposed

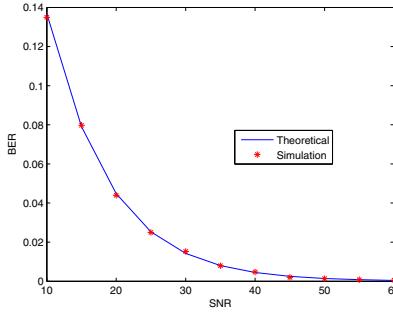


Fig. 3. Ideal relationship between SNR and BER

algorithm is theoretically valid. In reality, the relationship deviates from the ideal situation. It is reported in [9] that the errors can be on the same order as the SNR values. It is not known yet to which extent the same methodology could work for images. In the next section we show experiment results with real data.

4 Experiment Results

The proposed algorithm has been tested using 900 gray-scale natural-scene photos. The results can also extend to color images. Their content includes architecture, art, industry, mechanics, animals, humanoids, landscapes, vehicles, etc. The image size varies from 960×1280 to 1920×2560 pixels. We focus on the algorithm's capability of distortion estimation. Three kinds of distortions are considered – the JPEG compression, the JPEG2000 compression, and additive white Gaussian noise. They are commonly encountered in practice. The distortion is measured by SNR. However, we use a slightly modified definition of SNR – the mean value of the original image is subtracted from both the original signal x and the distorted signal y before calculation:

$$\text{SNR}[x, y] = \text{SNR}[x - \text{mean}(x), y - \text{mean}(x)] . \quad (11)$$

In this way, we meet one of the assumptions that the input signal is zero-mean; the noise signal $y - x$ is still the same. This definition allows better estimation accuracy. In the following, we always refer to the zero-mean SNR. We evaluate the estimation performance by the error ratio

$$\text{error ratio} = \frac{|\text{SNR}_{\text{real}} - \text{SNR}_{\text{estimated}}|}{\text{SNR}_{\text{real}}} \times 100\% . \quad (12)$$

The SNR values are always expressed in dB.

For each original image in the test database, a few distorted versions are generated by the operations listed in Table 1. For each pair of original and distorted images, their hash values are computed and compared; the real SNR is compared with the estimated SNR, which is computed by inputting the BER

Table 1. Distortions to be estimated

Distortion	Parameter range (step)	Error ratio	Improved error ratio
JPEG	Quality Factor: 30–90 (10)	28.64%	3.96%
JPEG2000	Compression ratio: 0.3–0.9 (0.1)	22.33%	2.70%
AWGN	SNR: 20 dB–50 dB (5)	31.56%	17.04%

to (10). Figure 4 shows the SNR values versus the estimated ones. The x-axis is the real SNR, and the y-axis is the estimated SNR. Ideally, the estimated values (red crosses) should lie on the (blue dash-dotted) diagonal line. The spread of the “clouds” indicates that the estimation errors are relatively large. The average error ratios (22.33%–31.56%) are listed in the 3rd column of Table 1. They probably restrict the application to the classification among a few SNR levels.

The errors are essentially due to the non-Gaussianity and the dependence between signal and noise. Large estimation variance was also observed in [9]. The authors discussed some possible ways to reduce the variance, such as discarding unreliable hash bits, averaging across space/time. These approaches may not be suitable for images due to the limited data. In the next section, we propose our own method to improve estimation accuracy.

5 Improve Estimation Accuracy

The large estimation errors in the previous tests show that the assumptions in Section 3.1 do not hold for natural-scene images. Equation (9) is not exact in practice. Therefore, a better model is required. However, for the same SNR, both the BER and the estimated SNR exhibit large variance. That means a single model (curve) is not sufficient to describe the SNR-BER relationship for all images. Instead, a model should be derived for each image.

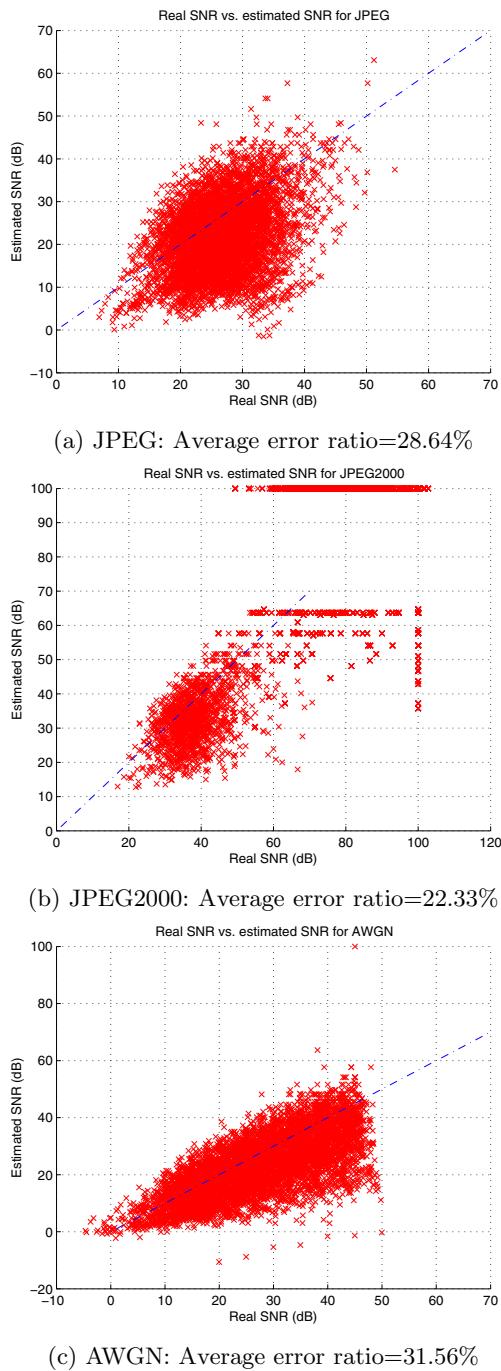
We assume that in practice (9) requires a correction term C :

$$P'_e = \frac{1}{\pi} \arctan \left(\sqrt{\frac{1}{\text{SNR}^2} + \frac{2}{\text{SNR}}} + C \right). \quad (13)$$

Correspondingly, (10) becomes:

$$\text{SNR} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, \quad (a = \tan(\pi P_e)^2 - C, b = -2, c = -1). \quad (14)$$

When estimating the SNR, best accuracy can be achieved if the correction term is available. However, storing all correction terms is certainly not a good option. A much more practical way is to predict the correction term from existing knowledge. Denote the SNR value estimated without any correction by SNR_0 . By observing the correction terms for all the distorted images in the previous section, we find that, for one image, the correction term for a specific SNR is

**Fig. 4.** Real SNR vs. estimated SNR

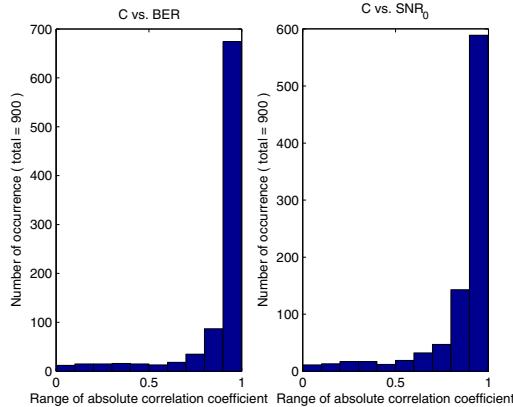


Fig. 5. Histogram of absolute correlation coefficients 1) between the correction term and the BER; 2) between the correction term and the SNR_0 for JPEG

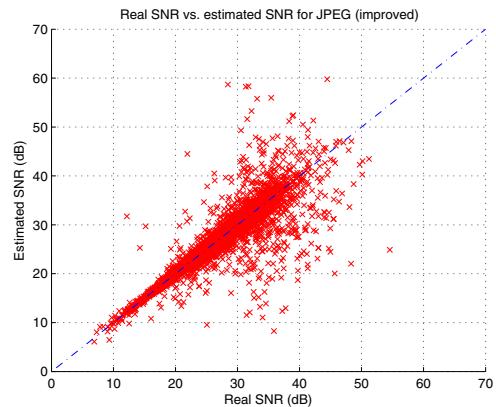
likely to be correlated to the corresponding BER and SNR_0 . Recall that, for each kind of distortion, 7 distorted images are generated for each original image. The vector of correction terms exhibits high correlation with the corresponding BER and SNR_0 vectors. Figure 5 shows the histogram of the absolute values of the correlation coefficients for JPEG. Most absolute correlation coefficients are close to 1. We observe the same situation for JPEG2000 and AWGN.

Based on the above observation, we assume there is a linear relationship between the correction term and the BER, SNR_0 values for a certain image:

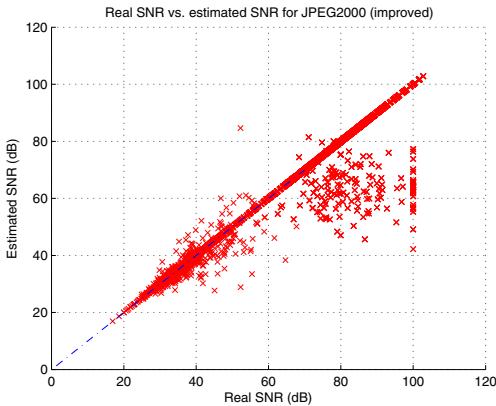
$$C = a \cdot \text{BER} + b \cdot \text{SNR}_0 + c . \quad (15)$$

Knowing C , BER and SNR_0 , the weights a, b, c can be derived by solving linear equations. We use the data in the previous experiment and compute a, b, c for each original image by the least-square method. The experiment is then run once again. The SNR estimation now consists of three steps: 1) compute SNR_0 by (9); 2) compute the correction term by (15); 3) compute SNR by (14). Figure 6 shows the new estimation results. The clouds are much more compact. The average error ratios are listed in the last column of Table 1. There is significant improvement for all cases. The error ratios for JPEG and JPEG2000 are reduced to less than 4%. The error ratio for AWGN is almost halved. Therefore, the proposed method is effective.

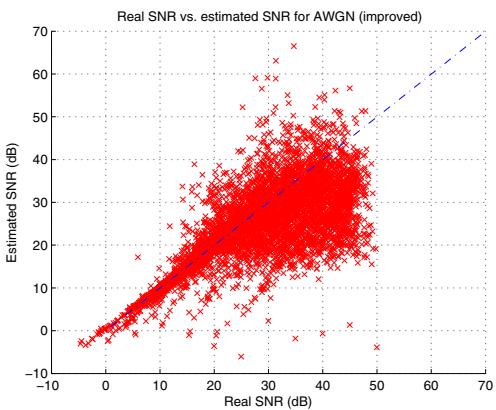
In the above approach, we only used one correction term. More correction terms are likely to provide better estimation. The price to pay is that we need to store the model parameters a, b, c for each original image, and their derivation requires some training data. There is a trade-off between accuracy, storage, and pre-computing.



(a) JPEG: Average error ratio=3.96%.



(b) JPEG2000: Average error ratio=2.70%.



(c) AWGN: Average error ratio=17.04%.

Fig. 6. Real SNR vs. estimated SNR (improved)

6 Conclusion

In this work, we explore a new application offered by image hashing techniques. By comparing the hash values of original images and their compressed versions, we are able to estimate the distortion level. A particular image hash algorithm is proposed for this application. The distortion level is measured by the signal to noise ratio (SNR). It is estimated from the bit error rate (BER) of hash values. The estimation performance is evaluated by experiments. The JPEG, JPEG2000 compression, and additive white Gaussian noise are considered. We show that a theoretical model does not work well in practice. The average error ratios are 28.64%, 22.33%, and 31.56% for the three kinds of distortion respectively. In order to improve estimation accuracy, we introduce a correction term in the theoretical model. We find that the correction term is highly correlated to the BER and the uncorrected SNR. Therefore it can be predicted using a linear model. A new estimation procedure is defined accordingly. New experiment results are much improved. The error ratios are reduced to 3.96%, 2.70%, and 17.04%.

References

1. Schneider, M., Chang, S.F.: A robust content based digital signature for image authentication. In: Proc. of International Conference on Image Processing (ICIP 1996), vol. 3, pp. 227–230 (1996)
2. Fridrich, J.: Robust bit extraction from images. In: Proc. of IEEE International Conference on Multimedia Computing and Systems, vol. 2, pp. 536–540 (1999)
3. Swaminathan, A., Mao, Y., Wu, M.: Robust and secure image hashing. *IEEE Transactions on Information Forensics and Security* 1(2), 215–230 (2006)
4. Schneier, B.: *Applied Cryptography: Protocols, Algorithms, and Source Code in C*, 2nd edn. John Wiley & Sons, Chichester (1996)
5. Menezes, A., van Oorschot, P., Vanstone, S.: *Handbook of Applied Cryptography*. CRC Press, Boca Raton (1996)
6. Roy, S., Sun, Q.: Robust hash for detecting and localizing image tampering. In: Proc. of International Conference on Image Processing, vol. 6, pp. 117–120 (2007)
7. Tagliasacchi, M., Valenzise, G., Tubaro, S.: Hash-based identification of sparse image tampering. *IEEE Transactions on Image Processing* 18(11), 2491–2504 (2009)
8. Lu, W., Varna, A., Wu, M.: Forensic hash for multimedia information. In: Proc. of SPIE Media Forensics and Security Conference (2010)
9. Doets, P.J.O., Lagendijk, R.L.: Distortion estimation in compressed music using only audio fingerprints. *IEEE Transactions on Information Forensics and Security* 16(2), 302–317 (2008)
10. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13(4), 600–612 (2004)
11. Brooks, A., Pappas, T.: Using structural similarity quality metrics to evaluate image compression techniques. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, April 15-20, vol. 1, pp. I-873–I-876 (2007)
12. Haitsma, J., Kalker, T.: A highly robust audio fingerprinting system. In: Proc. of 3rd International Conference on Music Information Retrieval, pp. 107–115 (October 2002)

Sewing Photos: Smooth Transition between Photos

Tzu-Hao Kuo¹, Chun-Yu Tsai¹, Kai-Yin Cheng¹, and Bing-Yu Chen²

National Taiwan University

{kakukogou,apfelpuff,keynes}@cmlab.csie.ntu.edu.tw, robin@ntu.edu.tw

Abstract. In this paper, a new smooth slideshow transition effect, *Sewing Photos*, is proposed while considering both of smooth content transition and smooth camera motion. Comparing to the traditional photo browsing and displaying work, which all focused on presenting splendid visual effects, *Sewing Photos* emphasizes on the smooth transition process between photos while taking the photo contents into account. Unlike splendid visual effects, the smooth transition tends to provide more comfortable watching experience and are good for a long-term photo displaying. To smooth the content transition, the system finds the similar parts between the photos first, and then decides what kind of camera operations should be applied according to the corresponding regions. After that, a virtual camera path will be generated according to the extracted Region of Interests (ROIs). To make the camera motion smoother, the camera path is generated as a cubic interpolation spline.

1 Introduction

People right now tend to take more photos at the same place [14], because of the advanced camera technologies and cheap storage. Hence, owing to the changed behavior, we do not only have more photos in one photo album, but also have more redundant information of a photo set.

To deal with the digital photos, traditional approaches usually utilize the extracted features from photos through a series of image analysis. The duplicated photos will be eliminated first before further processing [19]. Though some research work like Microsoft Photosynth¹ [18,17] and Photo Navigator [8] utilized the redundant information of a huge amount of photos to generate the 3D effects for photo browsing and presentation, they are not very suitable for creating a photo slideshow for casual usages.

According to the results from our field study, we found that though splendid slideshow visual effects like Animoto² can attract people's eyes, it is not suitable for playing repetitively and watching for a long duration. Hence, for different purposes, the needs will be quite contrary. However, the traditional researches for creating the slideshow visual effects usually focused more on the short-time

¹ <http://photosynth.net/>

² <http://animoto.com/>



Fig. 1. Three major types of normal users' photo taking behaviors

eye catching effects, rather than the needs for a long time playing. Besides advertising, there are many situations needed for a long-term playing including home photo displaying or repetitively playing photos in a digital photo frame. However, the traditional fade-in and fade-out effects do not take the photos' content into consideration, and just arbitrarily run through the photos, and the result is actually not really acceptable.

Through the observations of normal users' photo sets, we found that most of the photos can be categorized into following three types, as shown in Fig. ② (1) Global and detail: while taking pictures, users might zoom out to take an overview with the target object or scene, and zoom in to focus on some interesting details. (2) Pan to whole view: sometimes, the target object or scene is too large, and users might pan their cameras to capture the whole view, due to the limited view angle of the digital camera lens. (3) Same background: for some reasons, while users have events together, they might take pictures in turns in front of the landmarks or something representative.

Hence, if we can recover the camera operations, a simple but smooth transition might be achieved. However, rather than dealing with each case separately, a conceptual framework, *Buffer Region*, is proposed. Generally, we can find some similar parts between two photos. Then the found similar part(s) can be treated as the transition area between the two photos, where the virtual camera will pan or zoom through. Besides the smooth transition, to make the motion of camera smoother, the camera path is modeled as a cubic interpolation spline curve. Through this way, a simple but smooth visual effect can be produced.

2 Related Work

Generally, the photo slideshow displaying techniques can be categorized into 2D-based and 3D-based approaches.

In 2D-based photo slideshow, the most often seen and common used displaying effect is the Ken Burns effect³, which applies only zoom and pan operators to a

³ http://en.wikipedia.org/wiki/Ken_Burns_Effect/

given photo. Based on the Ken Burns effect, Microsoft Photo Story⁴ and their related research project, Photo2Video [9], developed a method to automatically extract ROI(s) in each photo and then the Ken Burns effect is applied to the camera motion according to the extracted ROI position(s) in the photo. Besides the Ken Burns effect, the other common used approach is the fade-in and fade-out effect. However, the normal fade-in/fade-out effect does not take the content into account, so the transition may not be smooth enough due to the difference between the two photos.

The ROI feature is also used by Tiling Slideshow [4], which clusters photos by time and content, and then arranges the photos within the same cluster into one frame with predefined tiling layouts. To browse several photos as whole, AutoCollage [15] combines each photo's ROI by blending them at boundaries. Though Tiling Slideshow and AutoCollage can show lots of information (photos) in one frame and can provide pretty exciting visual effects, they may not be suitable for a long-term displaying photo slideshow.

In addition to the 2D-based methods, some research work presents the photos in the 3D manner. Horry *et al.* proposed Tour Into the Picture (TIP) [7] which constructs a 3D space from one input photo by utilizing the proposed spidery mesh method. To construct the 3D space, Hoiem *et al.* proposed another approach, Automatic Photo Pop-up [6], which is a system that automatically constructs the 3D model from one single image and provides users a 3D geometrical view about the photo.

Instead of using a single photo, Hsieh *et al.* proposed Photo Navigator [8] which utilized the TIP method to create the 3D models of each photo and then connects them to generate a sequence of "walk-through" viewing path from one photo to another. Microsoft Photosynth¹ and their relative research projects, Photo Tourism [18][17], also utilized a number of photos to reconstruct a 3D space model and put the photos at the relative positions in it.

Though the above 3D-based methods aim to reconstruct the original 3D scene according to the extracted geometry features in the photos, they all need a huge amount of photos to construct the 3D space or need to take the pictures carefully. Hence, they may not be suitable for creating a photo slideshow for casual usages.

3 Field Study

3.1 Participants and Method

In order to know people's preference about slideshow effects, an interview was performed with 7 participants, who are 5 males and 2 females. For each participant, we let him/her watch a photo slideshow with some special transition effects and normal fade-in/fade-out effects. After watching the video, we asked them the following questions:

⁴ <http://www.microsoft.com/windowsxp/using/digitalphotography/photostory/default.mspx>

- What kinds of slideshow effects do you prefer?
- What will you do, if you will need to produce a photo slideshow?
- For different purposes, will the desired effects be differed?

3.2 Results and Findings

For the first question, most of them do not have specific preference. While asking the next two questions, the preference was revealed. For different purposes, people would use different kinds of slideshow effects, depending on the target audiences or the purposes of the photo slideshow. If the purpose is for advertising or the target audiences are strangers, they would prefer using the eye-catching effects. On the contrary, if the produced photo slideshow is for home video displaying, they would prefer using lightweight visual effects like the Ken Burns effect and simple fade-in/fade-out effect.

We quote some participants' explanations here: "Each different transition visual effect should have its own semantic meaning." "If the photos belong to the same topic, they should apply the same effect." (The mentioned effect here is the Ken Burns effect.) "The special transition effects can be used between two different topics as a hint to the audiences." "Too many special effects cause the video clutter and makes me feel bored finally, though it arouses my interest at first." "Special effects are some kinds of emphasis. You can't emphasize everything." "Rather than using special effects often, I prefer using lightweight effects, which makes the video watching comfortable."

The interview results can be concluded into three major findings. First, for different purposes and different audiences, the required transition visual effects are quite different. Second, compared with the special eye-catching effects, lightweight transition can be used all the time and would not cause any illness. Third, for the photos of the same topic, the transition should be smooth to make them as integral.

4 Framework of Buffer Region

4.1 Observation

Through the observation of users' photo sets, the photos can be generally classified into four major categories, which are *global and detail*, *pan to whole view*, *same background*, and *no-relationship*. The first three ones are shown in Fig. 11.

Global and detail - This kind of photos is always bound with camera's zoom operation. Users might zoom out to take an overview with the target object or scene, and zoom in to focus on some interesting details.

Pan to whole view - Sometimes, since the target object or scene is too large, in order to capture the whole view, users might need to pan their cameras. For this case, the usually used approach is to stitch the photos [19]. However, in our case, we just need to know the affine transformation between the two photos and generate a virtual camera path crossing through the two photos by registering them first.

Same background - While people go travel together, they might take pictures in turns in front of the landmarks or something representative. To demonstrate those photos, one possible approach is to use a photomontage method [2]. However, the content of the photo is altered. Hence, rather than altering the content, the similar parts between the two photos are detected and treated as the *Buffer Region* to let the virtual camera smoothly pass through.

No-relationship - For the photos have no above three relationships, they will be treated as no relationship.

4.2 Buffer Region

The core idea is to utilize the overlapped parts or similar regions as the smooth transition bridge between two photos. To achieve this goal, a similar system was proposed by Sivic *et al.* [16]. With the calculated GIST features [12], the system can automatically find the best "next" image for users to go forward in the virtual space endlessly. However, for people's casual photos without careful filtering, it may be hard to find so many similar photos.

Hence, to smoothly transit arbitrarily two photos, a so-called *Buffer Region* method is provided. As illustrated in the right lower two boxes in Fig. 2, the red rectangles are the extracted ROIs in both of purple and blue images and the green ones are the detected similar parts which are treated as the *Buffer Region*. Because the Rectangle C in the blue image is the most similar part with the Rectangle B in the purple one, therefore, while panning through the Rectangle B, the content inside the Rectangle B is smoothly replaced with the content inside the Rectangle C with alpha blending. Therefore, Rectangle B and Rectangle C are treated as the buffer regions while transiting from the purple image to the blue one. Through this way, the smooth visual effect can be generated.

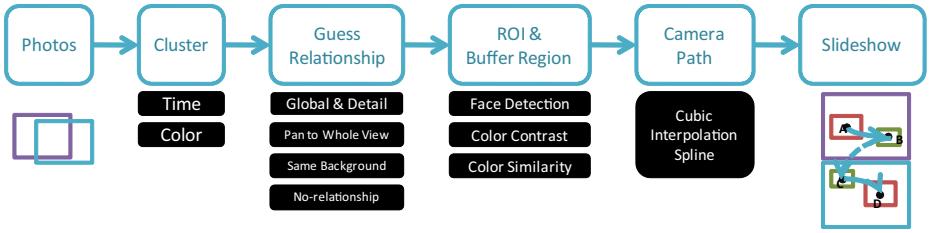
4.3 General Method

First, the aspect ratio should be the same with the original image. Second, the width and height of the region should not lower than W/n and H/n , where W and H are the width and height of the original image, and n is the scale factor. To make things simple, we let n be the number of power of two.

To calculate the similarity, we transfer the color space into $L^*a^*b^*$ model. Then the two norm operation is applied. To be general, each contribution can be weighted and the measurement of difference for one pixel can be written as:

$$\sqrt{\alpha \Delta L^2 + \beta(\Delta a^2 + \Delta b^2)}, \quad (1)$$

where $\alpha + \beta = 1$. Moreover, to remove the noise, before calculating, Gaussian smooth is applied first. To deal with the zoom case, the multi-scale approach is provided. When detecting the buffer region, the Image B will be downsized or enlarged with S levels to build a pyramid.

**Fig. 2.** System architecture

5 System

5.1 System Overview

The system structure is illustrated in Fig. 2. While inputting a set of photos, the system first clusters the photos based on the color similarity among them. After clustering the photos, the system matches each two adjacent photos in the same cluster by SIFT [11] features to compute the transition matrix of these two photos to guess the original camera operation. Due to the guessed camera operation, the relationship between the two photos can be determined. Then the ROI(s) and *Buffer Region* are extracted in each photo and the smooth camera path is calculated based on them. Finally, the photo slideshow is generated by playing through the ROIs and *Buffer Regions* with smooth transition.

5.2 Clustering Photos

Since photos in the same folder are usually captured consecutively at a certain location in a short period of time. Therefore, a two-step clustering method is used in this paper. The input photos are first clustered by the taken time, and then the clustered photos are further categorized according to the color similarity between each other.

Clustered by time - To cluster the input photos according to the taken time, we used the algorithm proposed in PhotoToc [13], which can dynamically decide the cutting threshold by using a sliding window.

Clustered by color - Though the photos in the same group clustered by the taken time, the content might be still varied. In our system, the color-based content similarity is used to categorize the photos in the same group, by assuming the color in the same theme is usually similar.

5.3 Guessing Camera Operations

To guess the original camera operations, the SIFT [11] features with RANSAC algorithm is used to match two adjacent images and then based on the matched corresponding points, the two images' transformation matrix can be calculated. If the two images are not similar enough, they will be regarded as the *no-relationship* case. Generally, the camera operation between the two photos in



Fig. 3. Camera move in the *global and detail* case

the same group containing both zoom and pan. Therefore, to discriminate the two types of *global and detail* and *pan to whole view*, the system will check whether one photo can be contained into another one or not.

5.4 Extracting ROIs and Buffer Regions

For each relationship, the ROIs and *Buffer Regions* are different. Generally, each photo itself will be defaultly treated as one ROI and will be put into the camera path. To extract the ROI(s) of each photo, the top-down and bottom-up approaches are both used [10].

For the top-down approach, the OpenCV face detector is used. For the bottom-up approach, the color contrast is taken as the measurement. First, the saliency map of the image is calculated by using the method proposed by Achanta *et al.* [1]. Then, the segmentation of the image is calculated by using the method proposed by Felzenszwalb and Huttenlocher [5]. Based on the segmented results, for each segmented region, the saliency scores are aggregated and the average score is calculated. If the average score of a segment is larger than a defined threshold, the segment will be treated as a candidate of the ROIs. The threshold is defined as two-fold of the average score of the whole saliency map here.

5.5 Calculating Camera Path

To calculating the camera path, the basic idea of Photo2Video [9] is adapted. However, two aspects are modified to meet our goal. First, we also model the areas of the ROIs into the spline calculation.

Second, the start and end points of the path inside a photo is bound by the two *Buffer Regions*, one for connecting the previous photo and the other for connecting the next one.

5.6 Generating Slideshow

To generate the smooth slideshow transition, the camera operations are simulated by the determined relationship between the adjacent two photos. Fig. 3 illustrates the *global and detail* case, where the Fig.s 3 (a) and (b) are two adjacent photos captured through a zoom camera operation. Fig. 3 (c) illustrates the zoom relation of the two photos.

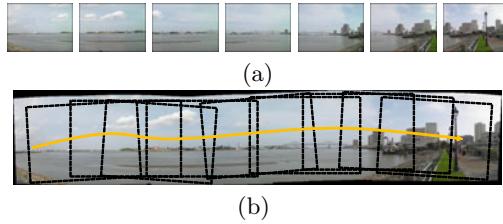


Fig. 4. Camera move in the *pan to whole view* case



Fig. 5. Camera move in the *same background* case

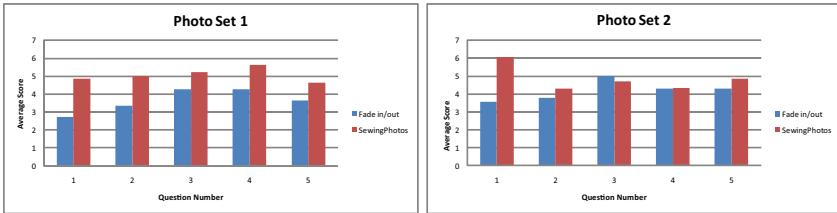
Fig. 4 is the illustration of the *pan to whole view* case. Fig. 4 (a) shows a series of consecutive photos captured by panning the camera from left to right. The system first tries to find the panorama [3] relationship of the set of photos under the pan operation and the result is as Fig. 4 (b). Then, according to the order of the input photos, the system applies a pan camera operation to replay the whole scene again.

Fig. 5 is the illustration of the *same background* case. However, the process is also used for the *no-relationship* case. After panning and zooming from the ROI region (red rectangle) to the *Buffer Region* (green rectangle) in Fig. 5 (a), the content inside the *Buffer Region* is smoothly replaced with the content inside the *Buffer Region* of Fig. 5 (b). Then, while finishing the transition, the camera continues panning and zooming to the ROI in Fig. 5 (b).

Besides the smooth transition, we also designed how long the camera should stay for each ROI according to the determined relationships. In the *global and detail* case, while the relationship is zooming in, it means that the next photo might be more important and more interested than the present one. Then, the camera should stay longer while playing the next photo comparing to the current one, and vice versa. While in the *pan to whole view* case, it means that the users care more about the whole view, not the detail in each transited photo. Then, the camera should not stay, but play through all the photos steadily. While in the *same background* case, of course, the users are interested in the foreground ROI(s). Therefore, it should stay on the foreground ROI(s) for a while. For the *no-relationship* case, the total display time for one photo is the same. Hence, if there are more than one ROI in a photo, the display time for each ROI is divided

Table 1. Evaluation Data

	Photo Set 1	Photo Set 2
Number of Photos	8	14
Composition	7 Pan 1 Same Background 10 Others	2 Zoom

**Fig. 6.** Evaluation Result

equally. Finally, while passing through the *Buffer Regions*, the camera should not stay. Through this way, the smooth slideshow video is generated.

6 Evaluation

Sewing Photos emphasizes on the smooth transition processes between photos, so we compared our result with the basic fade-in/fade-out transition effect. In the following evaluation, two photo sets are used. Table 1 listed the details of them, where Photo Set 1 is a panorama view of a river, and Photo Set 2 is captured in a reunion at a rabbit restaurant. In the following user testing, fifteen evaluators are invited. Their ages are from 22 to 28. The two slideshows of the three photo sets are played randomly.

We adopted the 7-points Likert-scale (1 is the worst, and 7 is the best) to design our evaluation scoring form. The evaluators were asked to score each video with the following five questions in different perspective:

1. **Fun:** Do they think it is an interesting presentation?
2. **Smoothness:** How do they think the smoothness at the transition?
3. **Experience:** How does the transition effect help them experience the trip?
4. **Camera Operation:** Do they think the applied camera operations suitable?
5. **Acceptance:** How about the willingness of adapting the transition effects?

The results are illustrated in Fig. 6, which show that the photo slideshows generated with *Sewing Photos* are generally better than the basic fade-in/fade-out transition effect, especially in Photo Set 1 which originally contains more zoom and pan relationships between the photos.

In Photo Set 2, the result of Question 4 (camera operation) is not significantly better, and even worse in Question 3 (experience). We think that it is because

Photo Set 2 mainly contains photos taken at the same background or with no specific camera operation, and such the advantages of adapting the original camera operations are not used, in the same time, the camera path applied by the system to those photos might not be suitable for this photo set which leads to the worse result in experience.

7 Conclusion

In this paper, we propose a novel method, *Sewing Photos*, which utilizes the original camera operations among the photos to generate a photo slideshow with smooth transition effects. The three major types of photo relationships in users' daily life photos are also defined. However, rather than dealing with each case individually, a general framework, *Buffer Region*, is proposed by utilizing the similar regions between the adjacent photos as the smooth transition tunnel. To smoothly play the photos, the motion of the camera is also taken into consideration. A cubic interpolation spline with the smallest curvature is calculated by using the extracted ROI(s) and *Buffer Regions* in one photo.

Though the alpha blending is good enough while transiting between two photos, a better method can be explored in the future by considering the human attention and camera direction. Though our system is treated as an engine to provide the smooth visual effects, to be an authoring tool, a user interface might be helpful, which can help the users to refine the results of the extracted ROIs and *Buffer Regions*, because some semantic ROIs are hard to be retrieved through low level contrast-based operation.

Concluding our work, the provided new technique is not going to replace the splendid visual effects in present the photo slideshow, but to enhance the traditional fade-in/fade-out effects to meet the needs. Nevertheless, the proposed smooth effect can provide more comfortable watching experience and are good for a long-term photo displaying.

Acknowledgments

This paper was partially supported by the National Science Council of Taiwan under NSC98-2622-E-002-001-CC2 and also by the Excellent Research Projects of the National Taiwan University under NTU98R0062-04.

References

1. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: Proceedings of the 2009 IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1597–1604 (2009)
2. Agarwala, A., Dontcheva, M., Agrawala, M., Drucker, S., Colburn, A., Curless, B., Salesin, D., Cohen, M.: Interactive digital photomontage. ACM Transactions on Graphics 23(3), 294–302 (2004); Proceedings SIGGRAPH 2004 Conference

3. Brown, M., Lowe, D.G.: Recognising panoramas. In: Proceedings of the 2003 IEEE International Conference on Computer Vision, vol. 2, pp. 1218–1225 (2003)
4. Chen, J.C., Chu, W.T., Kuo, J.H., Weng, C.Y., Wu, J.L.: Tiling slideshow. In: ACM Multimedia 2006 Conference Proceedings, pp. 25–34 (2006)
5. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision 59(2), 167–181 (2004)
6. Hoiem, D., Efros, A.A., Hebert, M.: Automatic photo pop-up. ACM Transactions on Graphics 24(3), 577–584 (2005); Proceedings SIGGRAPH 2005 Conference
7. Horry, Y., Anjyo, K.I., Arai, K.: Tour into the picture: using a spidery mesh interface to make animation from a single image. In: Proceedings ACM SIGGRAPH 1997 Conference, pp. 225–232 (1997)
8. Hsieh, C.C., Cheng, W.H., Chang, C.H., Chuang, Y.Y., Wu, J.L.: Photo navigator. In: Proceeding ACM Multimedia 2008 Conference, pp. 419–428 (2008)
9. Hua, X.S., Lu, L., Zhang, H.J.: Automatically converting photographic series into video. In: Proceedings ACM Multimedia 2004 Conference, pp. 708–715 (2004)
10. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998)
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
12. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
13. Platt, J.C., Czerwinski, M., Field, B.A.: PhotoTOC: Automatic clustering for browsing personal photographs. In: Proceedings of the 2003 IEEE Pacific Rim Conference on Multimedia, vol. 1, pp. 6–10 (2003)
14. Rodden, K., Wood, K.R.: How do people manage their digital photographs? In: Proceedings ACM CHI 2003 Conference, pp. 409–416 (2003)
15. Rother, C., Bordeaux, L., Hamadi, Y., Blake, A.: Autocollage. ACM Transactions on Graphics 25(3), 847–852 (2006); Proceedings SIGGRAPH 2006 Conference
16. Sivic, J., Kaneva, B., Torralba, A., Avidan, S., Freeman, W.T.: Creating and exploring a large photorealistic virtual space. In: Proceedings of the First IEEE Workshop on Internet Vision, pp. 1–8 (2008)
17. Snavely, N., Garg, R., Seitz, S.M., Szeliski, R.: Finding paths through the world's photos. ACM Transactions on Graphics 27(3), article no:15 (2008); Proceedings SIGGRAPH 2008 Conference
18. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. ACM Transactions on Graphics 25(3), 835–846 (2006); Proceedings SIGGRAPH 2006 Conference
19. Szeliski, R.: Image alignment and stitching: a tutorial. Foundations and Trends in Computer Graphics and Vision 2(1), 1–104 (2006)

Employing Aesthetic Principles for Automatic Photo Book Layout

Philipp Sandhaus¹, Mohammad Rabbath¹, and Susanne Boll²

¹ OFFIS – Institute for Information Technology, Oldenburg, Germany
firstname.lastname@offis.de

² Carl-von-Ossietzky Universität Oldenburg, Oldenburg, Germany
susanne.boll@informatik.uni-oldenburg.de

Abstract. Photos are a common way of preserving our personal memories. The visual souvenir of a personal event is often composed into a photo collage or the pages of a photo album. Today we find many tools to help users creating such compositions by different tools for authoring photo compositions. Some template-based approaches generate nice presentations, however, come mostly with limited design variations. Creating complex and fancy designs for, e.g., a personal photo book, still demands design and composition skills to achieve results that are really pleasing to the eye – skills which many users simply lack. Professional designers instead would follow general design principles such as spatial layout rules, symmetry, balance among the element as well color schemes and harmony. In this paper, we propose an approach to deliver principles of design and composition to the end user by embedding it into an automatic composition application. We identify and analyze common design and composition principles and transfer these to the automatic creation of pleasant photo compositions by employing genetic algorithms. In contrast to other approaches, we strictly base our design system on common design principles, consider additional media types besides text in the photo book and specifically take the content of photos into account. Our approach is both implemented in a web-based rich media application and a tool for the automatic transformation from blogs into photo books.

Keywords: photo composition, photo book, photo album, genetic algorithms, aesthetics, photo collage, multimedia presentation.

1 Introduction

Photos are a popular means to preserve the memory to important events in one's life. They document a baby's first developments, a vacation to an exotic destination, or the wedding of a happy couple. To remember and share these experiences, people tend to create visual compositions such as collages, calendars, or photo books. By the layout and style of the composition the users aim to create visually appealing presentations which should reflect his or her experience of the event captured in the photos. However, many users do not have the compositional and technical skills needed in a creative authoring process with today's commercial editing programs. A common approach to address this issue is to provide professionally designed templates as in commercially

available authoring tools¹². Presentations created this way usually lead to a pleasing result but often rather factual and uniform presentation.

Skilled users might exploit all the different options of professional graphics authoring tools and typically spend much time to manually achieve an if at all satisfying result. Depending on the user's abilities the results might be great but given the skills and the time the author needs to create the composition this is addressing only a few expert users. In a time in which the number of digital photos is exploding, companies in the field of photo finishing and photo services are searching for new products and business models to convert these photos into visually appealing digital presentations or physical products. We argue that the driving factor for success will be to "bring design to the people", i.e., to transfer knowledge from visual arts and design into a wizard-like authoring environment. We can literally hear the outcry by professional designers saying that layout and design will never be automatable and that the end result will never be meeting the standards of a manually and professionally curated presentation. Well, let's give it a try and see how far we can get for the end user.

In this paper, we present an approach for the automatic generation of photo compositions based on fundamental design and layout principles. We follow a generative approach in which the photo composition is created from a set of underlying rules. Each composition is unique and arranged according on the individual photo set. Existing approaches are either purely template-based or provide algorithms producing layouts that are usually not following any sophisticated, aesthetic design principles. We take a different approach and specifically take aesthetic rules as our starting point. Additionally, we explicitly consider not only photos but also other media types such as texts for the resulting layout and accommodate for the specific characteristics of these.

The visual layout of a photo only attributes to part of the overall quality and is embedded into processes which consists of other activities such as content selection and annotation [21]. We have addressed these activities in earlier works[45]. In this paper we assume that the content for a photo album has already been determined. We focus on the aspects of distributing these contents over the pages and generating pleasing page layouts.

We start by reviewing similar works in the field and comparing them with our approach in Section 2. We identify relevant design and layout rules from the literature and examples of professionally designed photo albums which are the basis for the development of our automatic layout system described in Section 3. These rules are the basis for our system for automatic photo book layout presented in Section 4. We describe two applications for our system in Section 5 before closing with a conclusion and outlook to future work.

2 Related Work

Most of the approaches for automatic layout and design of multimedia presentations aim at optimizing the page layout, e.g., based on minimal white space or maximization of the number of photos or regions of interest. The creation of photo collages

¹ <http://www.smilebooks.com>

² <http://www.apple.com/ilife/iphoto/>

presented by Grgensohn et al. [11] analyses photos for the region of interest and constructs stained-glass like photo collages from photos with faces. Other approaches aim to infer a tree-like structure on the photos and base their layout on these structure. For example, the approaches presented in [12][14] map the result of hierarchical clustering of photos directly to the spatial layout of the page. Others employ optimization techniques to minimize or maximize parameters such as the whitespace on the page or the occlusion of salient regions [23].

We also find approaches that derive clusters and importance of photos by content and context analysis and employ this information to select appropriate, pre-defined layout templates and placing the photos based on these [6][24]. AutoCollage [19] assembles a set of images on a canvas by alpha masks to hide joins between the different images and employs energy maps and region of interest detection for distributing the images over the page. In a template-based approach by Diakopoulos et al. [8], pre-designed templates are used which consist of cells for photos and annotations applied to these cells. The layout is filled by matching the metadata of photos to the annotations in the cells using an optimization algorithm. One of the few approaches that integrate design principles into the automatic presentation generation is presented by Lok et al. [16]. In their system, the authors introduce the concept of a WeightMap to solve the problem of visual balance for presentations. A work with a motivation similar to ours for automatic photo collage layout is the one proposed by Geigel et al. [10]. The authors try to mimic the artistic nature of the album layout process by employing genetic algorithms. A more recent work [9] presents an interesting authoring system for the selection, of photos for photo books and theme-based grouping, automatic background selection and automatic cropping. Following a similar goal, this approach however is based on a limited set of basic templates and does not consider images as backgrounds which, however, are very common in professionally designed photo books. Another recent work of the same group [13] presents an aesthetically-driven layout engine for the automatic generation of page-based presentations with pre-defined static content where pre-assigned areas can dynamically be filled with content whilst following aesthetic principles.

In conclusion, the approaches we find in the field provide methods which are algorithmically elegant but often do not sufficiently address the aesthetics of the generated results. We advance the work in the field in a number of ways: (1) We do not only consider photos as input to the system but also headings and text blocks, and we reflect their specific characteristics in the layout, taking into account that in the digital age photos often come with additional labels and descriptions. (2) We implement a page design that is strongly based on aesthetic principles stemming from common design rules and on the analysis of existing, professional designs. (3) We aim at creating compositions that are unique and reflect the features of just the individual media set.

3 Aesthetic Principles for Photo Books

One does not need to be a skilled designer to distinguish between an appealing and an unaesthetic photo book presentation. Some photo books instantly catch the viewer's eye and others are not really a pleasure to look at. Looking more closely at such photo books reveals certain patterns and applied rules regarding layout and design in the more appealing books. Many of these rules have been known for a very long time and are

employed by skilled artists and designers. In this section, we review selected rules for visual layout from the literature and discuss how they can be mapped to the automatic generation of photo books.

The rules and principles we consider for visual layout are mainly adopted from the works presented in [I2][I5]. Lidwell [I5] gives a good overview over universal design principles which are approved over many years by many designers Itten [I2] provides a good overview over different models for color combinations. In the following we further highlight two aspects of visual layout which has driven the design of our system. Besides these, we also reviewed a couple of professionally designed photo books.

3.1 Spatial Layout

Two of the central aspects of a layout are the size and position of the contained items. A well-known underlying principle creating the impression of balance and stability is the *golden ratio* (divine proportion) which describes a rule for the relation between two lengths: If a line ac is divided by a point b into the segments ab and bc the resulting sections follow the golden ratio if $bc/ab = ab/ac = (\sqrt{5} - 1)/2$. The golden ratio can be applied to many aspects in the photo composition such as the proportions of the size of different sub-areas or the position and relation of width and height of items on a page. Layouts the golden ratio are usually perceived as more balanced and appealing than layouts that do not. The golden ratio is well-known for the design of still images, e.g., by placing the horizon or the main object in a picture not in the middle of the image but at a point according to the golden ratio. Another principle for distributing objects and partitioning areas are different forms of *symmetry*. According to [I5], symmetry creates an impression of health and balance. Lidwell distinguishes between different kinds of symmetry: reflection, rotation, and translation symmetry. Reflection symmetry refers to the mirroring of an equivalent element around a central axis or mirror line. Rotation symmetry is referred to the rotation of equivalent elements around a common center. Translation symmetry refers to the location of equivalent elements with the same orientation and size in different areas of space. Translation symmetry is often combined with additional restrictions such as the placement of several elements along a line. These symmetry principles can, e.g., be seen in nature: A butterfly or a human's face exhibit reflection symmetry, a sunflower exhibits rotation symmetry in its petals around the center of the blossom.

3.2 Color Layout

When different colors are applied to the same design they usually lead to different perceived emotions [I5]. An important guideline is to limit the number of colors to the amount which can be processed at one glance. [I5] suggests about 5 colors for this. We can observe that some color layouts are perceived as more appealing than others. Looking more closely at such appealing color combinations one can observe a number of patterns. [I2] has structured such combinations of colors in different color schemes. One of the main attributes of a decent color layout is the limitation of colors. [I2]

suggests not to consider more than three and has structured such color combinations into six schemes. We employ these color schemes to determine appropriate color combinations for our automatic page layout system and to rate the overall color layout of a photo book page.

4 Automatic Photo Album Layout

The aesthetic principles for photo album layout discussed in the previous section form the basis for our proposed system for automatic layout of digital photo books. As input to our system we assume an ordered set of photos and text elements. Usually the order of the photos is derived from their time stamps but can also depend on the application. The text elements are assumed to be assigned to one photo or a series of photos. An origin of these kinds of media could be a travel blog with text and images from which one would like to create a photo book. But also other sources like photo management tools or social media platforms would come with photos and additional text in some kind of order. We provide descriptions of a few applications in Section 5. Based on the input media the system creates an automatic layout in five consecutive steps which are illustrated in Figure 1. The *Preprocessing* step analyzes and structures the input elements into different groups that form the basis for the different pages. These groups are then assigned to the different pages in the *Content Distribution* step. The creation of the layout begins with the determination of the *Background Layout* of the individual pages. The creation of the layout of the foreground is divided into two steps: For a High-Level Foreground Layout the pages are divided into separate areas according to certain layout principles, one for each element group. These areas are then laid out detailed in the *Detailed Foreground Layout* step.

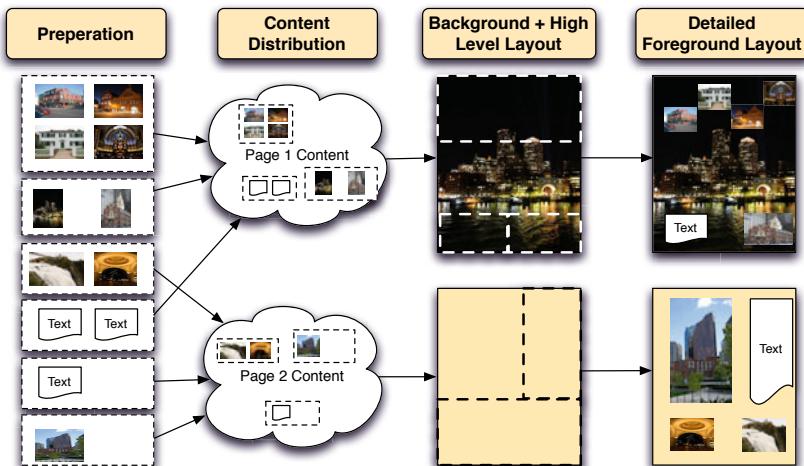


Fig. 1. System overview for our automatic photo album layout

4.1 Preprocessing

To be able to assign the different image and text items to the single pages, the content distribution step expects the elements to be structured in a certain way. This is accomplished in the preprocessing step. Basically two operations are performed on the input photos and text elements:

- In case no pre-defined order and structure is present in the photo set, they are ordered and clustered according to their time stamps. For this we employ the algorithm presented in [17]. By this we keep semantically close photos also together in the final photo book and thereby support the story-telling aspect of a photo album. The parameters of the clustering algorithm are chosen in a way that at most 5 photos form a group and the majority of groups only consists of one photo.
- Large text blocks are divided into several text blocks to also allow longer text passages in the photo book while avoiding cluttering the pages with text and maintaining a minimum threshold for the font size.

4.2 Content Distribution

The purpose of the content distribution step is to distribute the preprocessed elements over the photo book pages. For this we assume that the number of pages is predefined, which is reasonable in practical applications: For commercial photo book printing services, due to production limitations, usually only a fixed number of page numbers (e.g. 8, 16, 32, ...) and page sizes are offered. An optimal distribution of photos and text elements is defined by best meeting the following restrictions:

- R1 The predefined order of items should also be kept in the layout of the photo book.
- R2 The pages should be visually balanced, this means that roughly all pages should contain the same amount of items.
- R3 The text to photo ratio should be roughly the same on all pages. Therefore, the system should avoid to have pages only containing text or only photo items.
- R4 The content distribution should not extend the defined maximum of pages but at the same time avoid empty pages. This restriction stems from practical facts in some of our applications. Usually, when designing a printed photo book, one can only choose to increase the number of pages in multitudes of pages of, e.g. eight. A person who wants to order such an album usually wants to avoid having empty pages in the photo book but also wants to limit the number of pages as the price of the printed photo book increases with the number of pages.
- R5 Ensure that all page elements do really fit the page. Images should not be scaled down too much to ensure their visibility on the page and the font size of a the element should not be too tiny to be easily readable. This leads to minimum sizes both text and images.
- R6 The color layout of the page should be balanced, this means combinations of photos corresponding to one of the color schemes presented in Section 3 should be preferred.

The problem of content distribution can be seen as an optimization problem of which an optimal solution is that solution which best meets these partially competing conditions.

We opted to solve this problem with the help of a genetic algorithm. The result of the content distribution is an ordered set of pages each having assigned one or more groups of content elements.

4.3 Background Layout

After distributing all elements over the pages, the single pages are laid out individually. First for every page a background is determined as follows: An important principle regarding background vs. foreground of presentations is that the background should not distract from the foreground. We also have found out that the majority of photo books is designed by placing a suitable photo in the background. We therefore wanted to prefer layouts where a non-distracting photo is used as the background for a page and if no suitable photo can be determined, a color complementing the colors of the photos on the page should be taken.

In our system, distraction is modeled by two components: *Color* and *Saliency*. We assume that a photo with a lot of different colors distract the viewers eye more than photos only consisting of a few colors. We also assume that photos having a lot of salient regions like e.g. a photo showing a rough sea scenery do distract the viewer more than e.g. a photo with a very calm sea scenery. Thus, we have modeled these two aspects as follows:

- *Color*: To determine the degree of colorfulness of a photo we analyze the histogram of the photo in the HSV Color space. We only consider the histogram of the H component and count the bins which exceed a certain threshold value. The degree of colorfulness is then determined by the number of bins exceeding this value.
- *Saliency*: For every image a saliency map is determined. This map indicates regions in a photo which potentially catch the attention of the viewer. For this map we employ an extended version of the algorithm presented in [13]. The saliency map from the original algorithm is overlaid with a gaussian filter, which amplifies regions near the middle of the photo. This stems from our observations, that the important parts of an image are usually placed in the middle of the photos of lay photographers. Additionally, we employ the face detection algorithm of [22] to amplify regions containing faces. A score for the overall saliency of the image is determined by simply calculating the average saliency value of the resulting saliency map.

Photos exceeding a certain threshold value are automatically rejected as a background photo candidate. For the remaining candidates additionally the number of faces in each photo is determined. Then the photo with the least number of faces is chosen as the background photo. If there are more than one photo with the least number of faces the one is chosen with the least distraction rating. If there are no remaining photos a uniformly colored background is generated which best complements the main colors presents in the photos.

4.4 High-Level Foreground Layout

The design of the page's foreground is divided into two phases, a rough and a detailed layout process. In this first step the page area is divided into several rectangular areas,

one for each group. The spatial layout of these areas follows several, partly competing restrictions:

- RL1 All elements of one group have to fit in the assigned area.
- RL2 The pre-defined order of connected groups (text groups) should be reflected in the layout.
- RL3 Keep text items within their preferred spatial extension.
- RL4 Prefer overall layouts following principles of golden section and symmetry.
- RL5 Prefer aspect ratios following the golden ratio for text items.
- RL6 Avoid changing the aspect ratio of groups with one photo.
- RL7 Prefer keeping headings in the upper part of the page.
- RL8 Do not cover salient areas of the background.
- RL9 Items should be evenly distributed over the page and thus the visual weight should be placed in the middle of the page.

To meet these rules we again formulated the automatic design realized a genetic algorithm with a fitness function rating designs according to these rules. The generation of rough layouts is accomplished by recursively dividing the page into rectangular areas following principles of reflection symmetry and the golden ratio.

For this, we model the high-level page layout as a hierarchical graph with each inner leaf holding to values *hOrV* to decide if the remaining area is split horizontally or vertically and *splitAmount* to decide to which degree. The values of *splitAmount* are chosen according to the golden ratio. The leaves of the graph represent a sub-area on the page. Two examples of such layout graphs and the resulting layout is depicted in Figure 2.

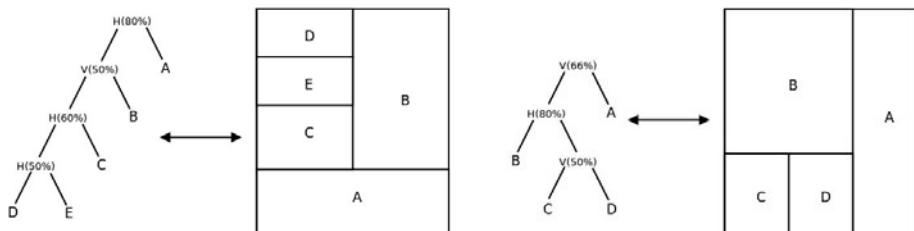


Fig. 2. Examples of resulting layouts according to recursive algorithm

The fitness function of the genetic algorithm evaluating the resulting layout consists of two parts. The first part consists of functions testing for mandatory conditions to the layout. If these conditions are not met, the corresponding chromosome is deleted from the population. These conditions consist of the above mentioned restrictions RL1-RL3. If these restrictions are met, the chromosome is rated according to the following definition (we also indicate in brackets [] which restriction is addressed):

- $r_1 = \frac{\#bord_{rat}}{\#bord}$: This determines the percentage of borders in the layout, which follow the golden section. Considering the nested nature of constructed layouts, it can happen, that borders do not follow this rule despite they are split according to the pre-defined slice primitives. This can also happen when considering aspect ratios of pages, which do not follow the golden section. [RL4]

- $r_2 = \frac{\#text_{gs}}{\#text}$: The percentage of text items on the page, which aspect ratios follow the golden section. [RL5]
- $r_3 = \frac{\sum_{i \in images} coveredArea_i}{\#images}$: The percentage to which the area can be covered by the image for single image groups. This favors layouts following the aspect ratio of the image and downgrades areas being degraded to long stripes or not following the image's aspect ratio. [RL6]
- $r_4 = \frac{\sum_{h \in headings} distanceToBottom_h}{\#headings}$: For every heading in the page it's distance to bottom of the page is determined favoring headings being in the upper part of the page. [RL7]
- $r_5 = salienceSum(freeArea)$: If a page consists of a background photo, one empty group is added to the page to allow keeping parts of the page free showing a very salient background. This rating determines the average salience value of the part of the background being covered with this empty group. If the page does not contain a background image, this part of the rating is discarded. [RL8]
- $r_6 = 1 - \frac{\sum_{e \in elements_page} \frac{e_x - width_page/2}{width_page} + \frac{e_y - height_page/2}{height_page}}{\#elements_page}$: The visual balance of the page. By e_x and e_y the center of an element group e is defined. The closer the average of all of these is placed to the center of the page, the higher the corresponding rating for the visual balance is. [RL9]

The overall rating function is defined as $r_{layout} = \sum_{i=1}^6 r_i$.

4.5 Detailed Foreground Layout

Giving a high level layout for a page, the different sub-areas are further laid out depending on the kind of their content. Areas containing only a **single photo** are filled entirely with this photo while keeping a small border. If the photo does not have the same aspect ratio as the assigned sub-area, it is cropped accordingly. For this the saliency map of the photo is determined and the photo is cropped in a way that the resulting photo corresponds to the aspect ratio of the sub-area while keeping the most salient regions of the photo.

Areas consisting of **text items** filled entirely with this text. Additionally we ensure, that the text is easily readable on the background. For this, as described in Section 4.3 we analyze the color histogram of the covered area of the background. If the background consists of only a few dominant colors (we chose 2), a color for the text is chosen which best complements these colors. If the background consists of too many dominant colors, the part of the background is overlaid with a translucent white area and the text color is set to black ensuring easy readability.

Areas consisting of **groups of photos** are laid out according to a set of pre-defined layout primitives. These are not templates but rather a set of visual rules to define the spatial relationships of a small set of photos. Figure 3 illustrates some of the layout primitives for our application which follow the principles of symmetry and the golden ratio. The figure shows only examples for a specific aspect ratio, the primitives are adjusted accordingly for different aspect ratios.

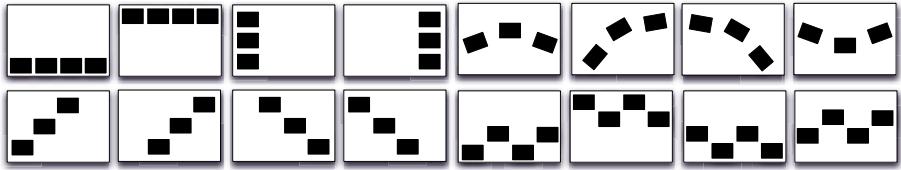


Fig. 3. Examples of layout primitives used for positioning photos inside a sub-area of a page

5 Application

One feature of our system is its ability to consider texts along with photos as input. Additionally, an already existing structure on these image and text items is also reflected in the resulting layout. This enables the system to be used in applications where these structures are present and textual content is associated together with the photos. Examples for this are e.g. travel blogs or photos hosted on social network platforms. Users who want to convert for example a photo into a photo book would appreciate if their valuable editing effort be maintained and reflected in the automatic layout. On the other side, our system is flexible enough to also provide sufficient results when this information is not available and can e.g. also work on raw photo sets. We deployed our layout system for the automatic generation of photo books from blogs [20] and social communication platforms [18]. Figure 4 shows the result of such an automatic transformation of a



Fig. 4. Example of two double-pages generated from a blog documenting a journey through Canada and the USA

travel blog documenting a trip through Canada and the USA into a digital photo book. Shown are four pages. On three pages a photo was chosen as the background and on the fourth page a matching, uniformly colored background was added. Headings from the original blog have resulted in headings and the entries' text and image contents have resulted in text and images on the pages laid out by our system [20].

6 Conclusion

We presented an approach for automatic layout of photo compositions which incorporates the knowledge about aesthetic design principles. We identified and analyzed rules

and principles of design in literature and from existing usage in professional photo album pages for their applicability in the domain of photo compositions. The discussed principles of layout and design were systematically integrated the design of our automatic layout system. We implemented our approach both in a web-based rich media application for the manual and automatic creation of photo compositions and a system for the automatic transformation of blogs and social media into photo books.

With our automatic design aid end users are now able to create layouts that are appealing with very low effort and limited design skills. The effectiveness of our system could be validated by an informal user study with 10 subjects which showed that the advantages of an automatic creation of appealing compositions from photos were appreciated by the users. The study supports the proposed kind of design support and shows that the resulting presentation are to the users' satisfaction. Although the subjects of our user study were very pleased with the results produced by our system we see potential for enhancement. In addition to the layout process we aim to support the initial selection of photos. Not only the design but also the selection of photos may follow aesthetics of photographs as proposed by Datta et al. [7] who aim to qualitatively distinguish between pictures of high and low aesthetic value.

References

1. Atkins, B.: Adaptive photo collection page layout. In: ICIP 2004, vol. 5, pp. 2897–2900 (October 2004)
2. Atkins, B.: Blocked recursive image composition. In: ACM MM 2008, pp. 821–824 (2008)
3. Balinsky, H.Y., Howes, J.R., Wiley, A.J.: Aesthetically-driven Layout Engine. In: Proceedings of the DocEng 2009, pp. 119–122. ACM Press, New York (2009)
4. Boll, S., Sandhaus, P., Scherp, A., Thieme, S.: Metaxa—context- and content-driven metadata enhancement for personal photo books. In: Cham, T.-J., Cai, J., Dorai, C., Rajan, D., Chua, T.-S., Chia, L.-T. (eds.) MMM 2007. LNCS, vol. 4351, pp. 332–343. Springer, Heidelberg (2006)
5. Boll, S., Sandhaus, P., Scherp, A., Westermann, U.: Semantics, content, and structure of many for the creation of personal photo albums. In: ACM Multimedia, p. 641. ACM Press, Augsburg (September 2007)
6. Chen, J.C., Chu, W.T., Kuo, J.H., Weng, C.Y., Wu, J.L.: Tiling slideshow. In: ACM MM 2006, pp. 25–34. ACM Press, New York (2006)
7. Datta, R., Li, J., Wang, J.Z.: Learning the Consensus on Visual Quality for Next-Generation Image Management. In: Proceedings of the ACM Multimedia Conference (September 2007), <http://infolab.stanford.edu/~wangz/project/imsearch/ALIP/ACMMM07A/>
8. Diakopoulos, N., Essa, I.: Mediating photo collage authoring. In: UIST 2005, pp. 183–186. ACM Press, New York (2005)
9. Gao, Y., et al.: MagicPhotobook: Designer inspired, User perfected Photo Albums. In: Proceedings of MM 2009, pp. 979–980. ACM Press, New York (2009)
10. Geigel, J., Loui, A.: Using genetic algorithms for album page layouts. IEEE MultiMedia 10(4) (2003)
11. Girsensohn, A., Chiu, P.: Stained glass photo collages. In: Proceedings of ACM Symposium on User Interface Software and Technology, pp. 13–14 (2004)
12. Itten, J.: The Art of Color: The Subjective Experience and Objective Rationale of Color. Wiley & Sons, Chichester (1997)

13. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(11), 1254–1259 (1998)
14. Kustanowitz, J., Shneiderman, B.: Meaningful presentations of photo libraries: rationale and applications of bi-level radial quantum layouts. In: *JCDL 2005*. ACM Press, New York (2005)
15. Lidwell, W., Holden, K., Butler, J.: *Universal Principles of Design*. Rockport Publishers (2003)
16. Lok, S., Feiner, S., Ngai, G.: Evaluation of visual balance for automated layout. In: *IUI 2004*, pp. 101–108. ACM Press, New York (2004)
17. Platt, J.C., Czerwinski, M., Field, B.A.: *PhotoTOC: Automatic Clustering for Browsing Personal Photographs*. Tech. rep., Microsoft Research (2002)
18. Rabbath, M., Sandhaus, P., Boll, S.: Automatic Creation of Photo Books from Stories in Social Media. In: *2nd SIGMM Workshop on Social Media (WSM 2010)*, Florenze, Italy (2010)
19. Rother, C., Bordeaux, L., Hamadi, Y., Blake, A.: Autocollage. In: *SIGGRAPH 2006*, pp. 847–852. ACM Press, New York (2006)
20. Sandhaus, P., Rabbath, M., Boll, S.: Blog2Book - Transforming Blogs into Photo Books Employing Aesthetic Principles. In: *ACM Multimedia 2010 - Technical Demonstrations*, Florenze, Italy (2010)
21. Sandhaus, P., Thieme, S., Boll, S.: Processes of photo book production. *Special Issue of Multimedia Systems Journal on Canonical Processes of Media Production* (2008)
22. Viola, P., Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition* (2001)
23. Wang, J., Quan, L., Sun, J., Tang, X., Shum, H.Y.: Picture collage. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 347–354 (2006)
24. Xiao, J., Zhang, X., Cheatle, P., Gao, Y., Atkins, C.B.: Mixed-initiative photo collage authoring. In: *ACM MM 2008*, pp. 509–518. ACM Press, New York (2008)

Video Event Retrieval from a Small Number of Examples Using Rough Set Theory

Kimiaki Shirahama¹, Yuta Matsuoka², and Kuniaki Uehara²

¹ Graduate School of Economics, Kobe University,
2-1, Rokkodai, Nada, Kobe, 657-8501, Japan

² Graduate School of System Informatics, Kobe University,
1-1, Rokkodai, Nada, Kobe, 657-8501, Japan

shirahama@econ.kobe-u.ac.jp, matuoka@ai.cs.scitec.kobe-u.ac.jp,
uehara@kobe-u.ac.jp

Abstract. In this paper, we develop an example-based event retrieval method which constructs a model for retrieving events of interest in a video archive, by using examples provided by a user. But, this is challenging because shots of an event are characterized by significantly different features, due to camera techniques, settings and so on. That is, the video archive contains a large variety of shots of the event, while the user can only provide a small number of examples. Considering this, we use “rough set theory” to capture various characteristics of the event. Specifically, by using rough set theory, we can extract classification rules which can correctly identify different subsets of positive examples. Furthermore, in order to extract a larger variety of classification rules, we incorporate “bagging” and “random subspace method” into rough set theory. Here, we define indiscernibility relations among examples based on outputs of classifiers, built on different subsets of examples and different subsets of feature dimensions. Experimental results on TRECVID 2009 video data validate the effectiveness of our example-based event retrieval method.

Keywords: Example-based event retrieval, Rough set theory, Bagging, Random subspace method, high-dimensional small sample size problem.

1 Introduction

Recently, there is a great demand to develop a method which can efficiently retrieve events of interest in a video archive. It is impractical or laborious to pre-index or prepare models for all possible events. Thus, we focus on example-based event retrieval which does not need any pre-indexing and model preparation, but only uses examples provided by a user.

Existing example-based event retrieval methods can be classified into two categories, “similarity-calculation approach [1][2]” and “model-construction approach [3][4]”. Similarity-calculation approach only uses “positive examples (p-examples)” where an interesting event is shown. But, meaningful events cannot be defined only by using p-examples. For example, for the event “a car moves”, if a p-example where a red car moves is provided, it may be more similar to a shot where a person

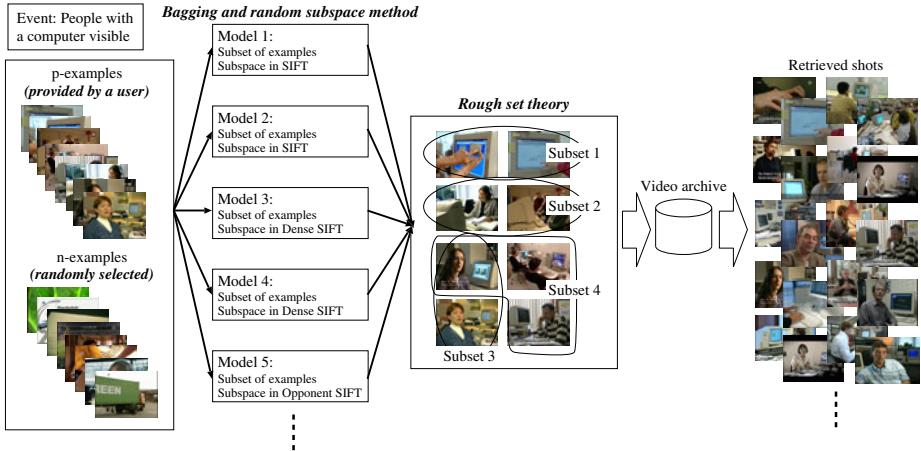


Fig. 1. An overview of our example-based event retrieval method

wears a red cloth than a shot where a white car moves. Like this, only by similarities, many irrelevant shots are inevitably retrieved.

On the other hand, model-construction approach uses both p-examples and “negative examples (n-examples)” where the event is not shown. Generally, p-examples are used to generalize a retrieval model, while n-examples are used to specialize the model [5]. More specifically, by adding a p-example, the model is generalized to include shots similar to the p-example into the retrieval result. But, it potentially retrieves many irrelevant shots (i.e. false positive). So, by adding an n-example, the model is specialized to exclude shots similar to the n-example from the retrieval result. In this way, the model can discriminate relevant and irrelevant shots to the event. Therefore, model-construction approach can achieve a much more accurate retrieval than similarity-calculation approach. In this paper, we concentrate on developing an example-based event retrieval method based on model-construction approach.

2 Addressed Problems

Fig. 1 shows an overview of our method. It must be noted that a user can only provide a small number of training examples by searching videos at hand or by using on-line video search engines. The video archive contains various shots of the event, which are taken by different camera techniques and in different settings. Thus, our main research objective is how to retrieve a large variety of shots of the event only by using a small number of examples.

To this end, we use “rough set theory” and “bagging and random subspace method”, as shown in the middle part of Fig. 1. Roughly speaking, we firstly build various models by using different subsets of examples (i.e. bagging) and different subsets of feature dimensions (i.e. random subspace method). Here, although these models can detect shots of the event with different characteristics, they

are not so accurate due to the insufficiency of examples and feature dimensions. Thus, we use rough set theory to combine the above models to accurately retrieve shots of the event. Below, we describe motivations and advantages of using rough set theory, bagging and random subspace method.

2.1 Large Variation of Features in the Same Event

Depending on camera techniques and settings, shots of the same event contain significantly different features. But, we can find that subsets of shots are taken by similar camera techniques and settings, and are characterized by similar features. For example, two shots in *Subset 1* in Fig. 11 are characterized by hands and computer monitors. Also, two shots in *Subset 3* are similar because each of them takes a person in a tight shot. Furthermore, three shots in *Subset 4* are similar because people appears with computers in complex backgrounds. Considering this subset property, we cannot define an event only by using a single model, because subsets are characterized by different combinations of features. For example, although a feature related to hands is important for *Subset 1*, it is unimportant for the other subsets. Thus, we separately represent each subset by a different model, and define the event as the union of such subsets.

To implement the above idea, we use “Rough Set Theory (RST)” which is a set-theoretic classification method for extracting rough descriptions of a class from imprecise (or noisy) data [6]. Specifically, RST firstly determines the indiscernibility relation between a p-example and an n-example, that is, whether they can be discerned with respect to available features. Then, by combining such indiscernibility relations based on the set theory, RST extracts classification rules called “decision rules”. Each decision rule can correctly identify a subset of p-examples, such as subsets in Fig. 11.

Note that a traditional RST can deal only with categorical features, where we can easily define the indiscernibility relation between two examples by examining whether they have the same value or not. But, in our example-based event retrieval, examples are represented by non-categorical high-dimensional features. For example, the bag-of-visual-words representation involves thousands of dimensions, each of which indicates the frequency of a visual word, defined on a local image feature like SIFT. Thus, when applying RST to our example-based event retrieval, the most important issue is how to define indiscernibility relations among examples, in other words, how to categorize a non-categorical high-dimensional feature space.

To categorize high-dimensional features, we use non-linear SVMs [7] because they are known as one of the best classifier for high-dimensional features. Specifically, the margin maximization offers a good generalization in a high-dimensional feature space. Also, the kernel representation is effective for a high-dimensional feature, since examples in the high-dimensional feature space are mapped into points in a much lower dimensional space. Therefore, in this paper, we develop RST where high-dimensional features are categorized by non-linear SVMs. Especially, we use SVMs with RBF kernel, since it is known as the most general kernel [8].

2.2 High-Dimensional, Small Sample Size Problem

As the number of dimensions of a feature increases, the number of examples needed to build a well generalized classifier exponentially increases. It is because we need to consider whether each combination of values in different dimensions characterizes a class or not. But, compared to a high-dimensional shot representation with more than 1,000 dimensions, a user can only provide a much smaller number of examples. This is called a “high-dimensional, small sample size problem”.

In our example-based event retrieval, the high-dimensional, small sample size problem can be exemplified by the following two typical cases. First, an insufficiency of p-examples causes overfitting. This means that retrieved shots match with feature dimensions, where p-examples are occasionally similar to each other. For example, for the event “people talk”, if all of p-examples are taken outdoors, shots only showing outdoor situations are preferred to shots where people talk indoors. Second, an insufficiency of n-examples causes overgeneralization. That is, retrieved shots match with feature dimensions where n-examples are occasionally similar to each other. As a result, in these dimensions, a large region of p-examples is estimated and includes many irrelevant shots. For example, for the above event, if all of n-examples show non-human objects, shots where people do any actions may be retrieved.

For the above high-dimensional, small sample size problem, we use “bagging [9]” to avoid overgeneralization. Specifically, we build several SVMs on different sets of randomly selected examples. And, by combining of these SVMs, we aim to extract a region which accurately characterizes shots of an event. Furthermore, we use “random subspace method [10]” to overcome overfitting. Specifically, by building several SVMs on different subsets of randomly selected feature dimensions, we can avoid emphasizing feature dimensions where p-examples are occasionally similar to each other.

Our method is similar to existing event retrieval methods [13,3] in the sense that bagging and random subspace method are used. But, our method is crucially different from [13,3], because we do not simply integrate SVMs by using the majority voting, but integrate SVMs by using RST. With respect to this, accuracies of SVMs significantly depend on selected examples and feature dimensions. If the majority voting is performed using all SVMs, the retrieval performance is clearly degraded by inaccurate SVMs. So, we use RST to extract decision rules as combinations of SVMs, which can accurately identify different subsets of p-examples. By using such decision rules, we perform the majority voting. Hence, our method can be considered as superior to the ones in [13,3]. Finally, although [14] proposes RST which categorizes high-dimensional features by classifiers, it uses neither bagging nor random subspace method.

3 Example-Based Event Retrieval Method

In this section, we describe our example-based event retrieval method. First of all, by using the color descriptor software [11], we extract the following 6 different

types of features from the keyframe of each shot: 1. *SIFT*, 2. *Opponent SIFT*, 3. *RBG SIFT*, 4. *Hue SIFT*, 5. *Color histogram* and 6. *Dense SIFT*. For the space limitation, we omit the explanation of these features (see [1] in more detail). We represent the above 6 types of features by using the bag-of-visual-words representation. Specifically, for each type of feature, we extract 1,000 visual words by clustering features at 200,000 interest points, sampled from keyframes in TRECVID 2009 development videos [2]. Thus, a shot is represented as a total 6,000-dimensional vector, that is, 1,000 dimensions per feature.

Based on the above high-dimensional representation, we explain rough set theory (RST) enhanced by bagging and random subspace method. Particularly, we firstly explain the traditional RST [3], where we assume that high-dimensional features are already transformed into categorical features. Then, we present how to categorize high-dimensional features by using SVMs. Finally, we incorporate bagging and random subspace method into RST.

Given p-examples and n-examples for an event, we use RST to extract decision rules for discriminating shots of an interesting event from all other shots. Let p_i and n_j be i -th p-example ($1 \leq i \leq M$) and j -th n-example ($1 \leq j \leq N$), respectively. a_k indicates k -th feature ($1 \leq k \leq K$), where $a_k(p_i)$ and $a_k(n_j)$ represent categorical features of p_i and n_j for a_k , respectively. As shown in Fig. 2 (a), RST represents the above kind of p-examples and n-examples in the form of table, called “decision table”. Here, each row represents an example. The rightmost column indicates whether an example is positive (“P”) or negative (“N”), while the other columns indicate features. Like this, the decision table provides available information for discriminating between p-examples and n-examples.

a)	a_1	a_2	a_3	a_4	a_5	a_6	Class
p_1	$a_1(p_1)$	$a_2(p_1)$	$a_3(p_1)$	$a_4(p_1)$	$a_5(p_1)$	$a_6(p_1)$	P
p_2	$a_1(p_2)$	$a_2(p_2)$	$a_3(p_2)$	$a_4(p_2)$	$a_5(p_2)$	$a_6(p_2)$	P
n_1	$a_1(n_1)$	$a_2(n_1)$	$a_3(n_1)$	$a_4(n_1)$	$a_5(n_1)$	$a_6(n_1)$	N
n_2	$a_1(n_2)$	$a_2(n_2)$	$a_3(n_2)$	$a_4(n_2)$	$a_5(n_2)$	$a_6(n_2)$	N

b)

IF $a_1=a_1(p_1)$, THEN Class = P

IF $a_2=a_2(p_1)$ AND $a_3=a_3(p_1)$, THEN Class = P

IF $a_2=a_2(p_1)$ AND $a_5=a_5(p_1)$, THEN Class = P

Fig. 2. Example of a decision table (a) and decision rules (b) in RST

First, in order to define the indiscernibility relation between each pair of p_i and n_j , RST extracts “discriminative features” which are useful for discriminating them. Assuming that each feature is categorical, the set of discriminative features $f_{i,j}$ between p_i and n_j can be represented as follows:

$$f_{i,j} = \{a_k | a_k(p_i) \neq a_k(n_j)\} \quad (1)$$

$f_{i,j}$ means that when at least one feature in $f_{i,j}$ is used, p_i can be discriminated from n_j . Fig. 2 (a) illustrates the process of extracting discriminative features

$f_{1,1}$ between p_1 and n_1 , and $f_{1,2}$ between p_1 and n_2 . As a result, p_1 and n_1 can be discriminated by using one of $\{a_1, a_3, a_5\}$. Also, p_1 can be discriminated from n_2 by one of $\{a_1, a_2\}$.

Next, we extract combinations of features which are needed for discriminating p_i from all n-examples. This is achieved by using at least one feature in $f_{i,j}$ for all n-examples. That is, we take a conjunction of $\vee f_{i,j}$ as follows:

$$df_i = \wedge \{\vee c_{i,j} \mid 1 \leq j \leq N\} \quad (2)$$

For example, in Fig. 2 (a), df_1 is computed as $(a_1 \vee a_3 \vee a_5) \wedge (a_1 \vee a_2)$. This is simplified as $df_1^* = (a_1) \vee (a_2 \wedge a_3) \vee (a_2 \wedge a_5)$ ¹. That is, p_1 can be discriminated from all n-examples n_1 and n_2 , by using a_1 , the set of a_2 and a_3 or the set of a_2 and a_5 . Like this, each conjunction term in df_i^* represents a “reduct” which is a minimal set of features needed to discriminate p_1 from all n-examples.

From each reduct, we construct a decision rule in the form of *IF-THEN* rule. For example, from the above three reducts, we can construct decision rules shown in Fig. 2 (b). Here, each decision rule examines whether a test example and p_1 have the same values for features in the conditional part. Note that each decision rule in Fig. 2 (b) characterizes a subset consisting only of one p-example p_1 . By gathering decision rules extracted for all p-examples, we can obtain decision rules which characterize subsets of multiple p-examples. Finally, based on majority voting of decision rules, we retrieve shots which match with many decision rules.

Now, we discuss how to obtain categorical values for high-dimensional features by using SVMs. In one sentence, we build an SVM on each high-dimensional feature, and regard its outputs as categorical features in RST. This is exemplified by the decision table in Fig. 3 (a). Here, an SVM is built on each of 6 features, and each example is represented as a vector of SVMs’ outputs.

As shown in Fig. 3 (a), it is difficult to accurately discriminate between p-examples and n-examples only by using a single SVM. So, by using rough set theory, we extract reducts as minimal sets of SVMs which are needed to correctly discriminate a p-examples from all n-examples. A decision rule constructed from such a reduct examines whether a test example is classified as positive by multiple SVMs. For example, Fig. 3 (b) shows the decision rule which examines a test example by using two SVMs on *SIFT* and *Hue SIFT*.

However, due to the high-dimensional, small sample size problem, SVMs built on all examples and all dimensions may be overfit or overgeneralized (see section 2.2). And, decision rules as combinations of such SVMs become meaningless. Thus, we incorporate bagging [9] and random subspace method [10] into RST. Fig. 4 illustrates a decision table of RST extended by bagging and random subspace method. Here, *SVM1*, *SVM2* and *SVM3* are built on different subsets of p-examples and n-examples and different subsets of dimensions in *SIFT* feature. Also, *SVM4* is built on a different subset of examples and a different subset of dimensions in *Opponent SIFT* feature. By using RST, we combine the above

¹ This simplification is achieved by using the distributive law $A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C)$ and the absorption law $A \vee (A \wedge B) = A$.

a)

	SVM _{SI}	SVM _{O-SI}	SVM _{R-SI}	SVM _{H-SI}	SVM _{R-HI}	SVM _{D-SI}	Class
p ₁	+1	-1	-1	+1	+1	-1	P
p ₂	-1	+1	+1	-1	-1	-1	P
n ₁	+1	+1	-1	-1	+1	+1	N
n ₂	-1	-1	+1	+1	-1	+1	N

b)

IF SVM_{SI} = +1 **AND** SVM_{H-SI} = +1
, THEN Class = P

- SVM_{SI}: SVM on SIFT
- SVM_{O-SI} : SVM on Opponent SIFT
- SVM_{R-SI} : SVM on RGB SIFT
- SVM_{H-SI} : SVM on Hue SIFT
- SVM_{R-HI} : SVM on RGB Histogram
- SVM_{D-SI} : SVM on Dense SIFT

Fig. 3. Example of a decision table (a) and decision rules (b) in RST using SVMs

	SVM1	SVM2	SVM3	SVM4	SVM5	Class
p ₁	+1	+1	-1	+1	-1	P
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
p _M	-1	+1	+1	+1	-1	P
n ₁	-1	-1	-1	+1	-1	N
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n _N	-1	-1	-1	-1	+1	N

Fig. 4. A decision table in RST extended by bagging and random subspace method

kind of SVMs into decision rules, so that we can overcome overfitting and overgeneralization.

Note that other than avoiding overfitting and overgeneralization, building a larger number of SVMs leads to extracting a larger variety of decision rules. With respect to this, [13] describes that when only a small number of examples are available, an SVM is “unstable” where its classification result significantly change depending on used examples. Therefore, decision rules as combinations of such unstable SVMs can cover various shots of an event, even in the case of high-dimensional, small sample size.

4 Experimental Results

In this section, we test our method on TRECVID 2009 video data, which consists of 219 development videos (36, 106 shots) and 619 test videos (97, 150 shots) [12]. In particular, we evaluate our method on the following 5 events: *Event 1*: A view of one or more tall buildings and the top story visible, *Event 2*: Something burning with flames visible, *Event 3*: One or more people, each at a table or

desk with a computer visible, *Event 4*: an airplane or helicopter on the ground, seen from outside, *Event 5*: One or more people, each sitting in a chair, talking.

Each event is retrieved as follows. First, we manually collect p-examples from development videos. Then, by randomly selecting shots in development videos, we collect 5 times larger number of n-examples than the number of p-examples. This approach is appropriate from the probabilistic perspective, because the number of shots relevant to the event is very small. Next, we run our example-based event retrieval method to retrieve shots of the event in test videos. Here, we use two libraries, LIBSVM [8] and ROSETTA [6] for SVM learning and the reduct extraction in RST, respectively. SVM parameters are determined by 3-fold cross validation. Finally, we evaluate the retrieval performance as the number of relevant shots within 1,000 retrieved shots.

In order to examine the effectiveness of RST enhanced by bagging and random subspace method, we compare the following 4 types of retrieval:

1. *Baseline*: We simply build an SVM on each of 6 features, and search test videos by using the SVM. Then, from SVMs on all features, we manually select the SVM which yields the best result. Note that *Baseline* is favored, because the best feature for retrieving an event is unknown in advance.
2. *RST_only*: We run RST which uses neither bagging nor random subspace method. That is, one SVM is built on each feature by using all examples and all feature dimensions.
3. *RST+BG*: We run RST only by using bagging. In particular, we build three SVMs on each feature by using different subsets of examples and all feature dimensions. Each subset is made by randomly sampling 75% of examples.
4. *RST+BG+RS*: We run RST which uses both bagging and random subspace method. In particular, we build 10 SVMs on each feature by using different subsets of examples and different subsets of feature dimensions. Each subset of examples is made by randomly sampling 75% of examples, while each subset of feature dimensions is made by randomly sampling 50% of feature dimensions.

Fig. 5 shows performances of the above 4 types of retrieval. For each event, the number in parentheses represents the number of p-examples. Also, since p-examples, n-examples and feature dimensions randomly change in our method, Fig. 5 represents the average number of correct shots in 10 retrieval results.

As can be seen from Fig. 5, *RST_only* is not so effective because its performance for *Event 1*, *Event 3* and *Event 4* is similar to the one of *Baseline*. But, both of *RST+BG* and *RST+BG+RS* outperform *Baseline* for all events. This means that in *RST_only*, only from 6 SVMs (one SVM built on each feature), RST cannot extract decision rules which characterize various shots of an event. But, when the variety of SVMs is increased by bagging and random subspace method, RST can extract decision rules for characterizing various shots of the event. This result not only indicates the effectiveness of RST for covering various shots of an event, but also indicates the effectiveness of combining RST with bagging and random subspace method.

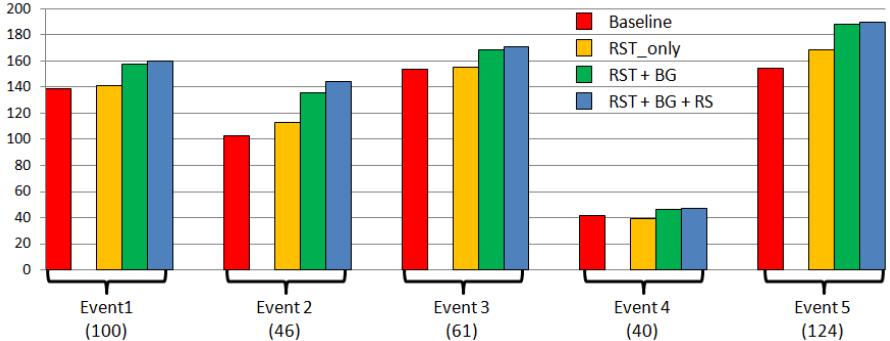


Fig. 5. Performance comparison among *Baseline*, *RST_only*, *RST+BG* and *RST+BG+RS*

Next, we examine whether our method can work when a small number of p-examples are available. Fig. 6 presents how to change retrieval performance of *Baseline*, *RST+BG* and *RST+BG+RS* depending on numbers of available p-examples. Here, we evaluate the performance as the average of 10 retrieval results.

For *Event 2* and *Event 3* in Fig. 5, *RST+BG* and *RST+BG+RS* always outperform *Baseline*. But, for the other events, when the number of available p-examples is very small like 10 or 20, *RST+BG* and *RST+BG+RS* are outperformed by *Baseline*. It can be considered that in such a case, most of SVMs built by bagging and random subspace method are inaccurate. So, by combining such inaccurate SVMs, we cannot extract useful decision rules. Overall, we can say that when more than 30 p-examples are available, *RST+BG* and *RST+BG+RS* can retrieve a larger number of relevant shots than *Baseline*.

From the above result, collecting 30 p-examples is easy for frequently occurring events such as *Event 1* and *Event 5*, but it may be difficult for rarely occurring events such as *Event 2* and *Event 4*. With respect to this, we are currently developing two types of methods for getting a sufficient number of p-examples. The one type of method utilizes image/video search engines on the web like Flickr and YouTube. The second type method creates “pseudo” p-examples by synthesizing user’s action, 3DCG and a background image using virtual reality technique.

Finally, we compare our method to state-of-the-art event retrieval methods, especially, methods submitted to the fully automatic category in TRECVID 2009 search task [12]. Overall, the performance of *RST+BG+RS* in Fig. 5 is ranked in about the top quartile. Specifically, among 88 retrieval results for *Event 1*, the average performance (0.0950) and the maximum performance (0.1068) are ranked at 16-th and 14-th positions, respectively. Also, for *Event 2*, the average performance (0.1918) and the maximum one (0.2027) are ranked at 12-th and 6-th positions, respectively. Note that almost all of methods in the top quartile adopt concept-based approaches. Here, in order to recognize the presence or absence of each concept in a shot, concept detectors are built in advance by using more than ten thousands of training examples. On the other hand, our method

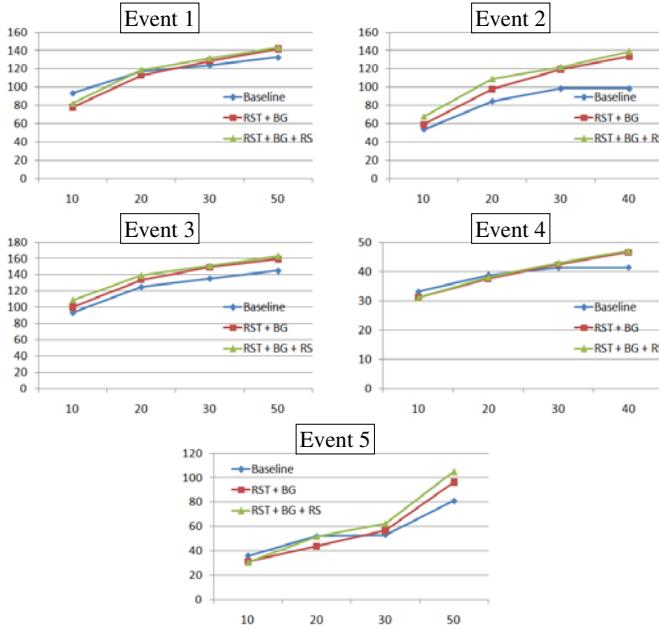


Fig. 6. Retrieval performances in different numbers of available p-examples

only uses examples that a user extemporarily provides. Therefore, we can say that our method is a very effective method which needs no model preparation like building of concept detectors.

5 Conclusion and Future Works

In this paper, we proposed an example-based event retrieval method which can retrieve a large variety of shots of an interesting event only from a small number of examples. To this end, we develop RST enhanced by bagging and random subspace method. Specifically, we firstly define indiscernibility relations among examples on high-dimensional features, by using SVMs built on different subsets of examples and on different subsets of feature dimensions. Then, based on these indiscernibility relations, we conduct RST to extract decision rules as combinations of SVMs, which can correctly identify different subsets of p-examples. Experimental results show that when more than 30 p-examples are available, our method successfully covers various shots relevant to an event. Also, we show that the performance of our method approaches to those of methods, which use models pre-constructed from a huge number of training examples.

References

- Peng, Y., Ngo, C.-W.: EMD-Based Video Clip Retrieval by Many-to-Many Matching. In: Leow, W.-K., Lew, M., Chua, T.-S., Ma, W.-Y., Chaisorn, L., Bakker, E.M. (eds.) CIVR 2005. LNCS, vol. 3568, pp. 71–81. Springer, Heidelberg (2005)

2. Kashino, K., Kurozumi, T., Murase, H.: A Quick Search Method for Audio and Video Signals based on Histogram Pruning. *IEEE Transactions on Multimedia* 5(3), 348–357 (2003)
3. Natsev, A., Naphade, M., Tešić, J.: Learning the Semantics of Multimedia Queries and Concepts from a Small Number of Examples. In: Proc. of ACM MM 2005, pp. 598–607 (2005)
4. Shirahama, K., Sugihara, C., Uehara, K.: Query-based Video Event Definition Using Rough Set Theory and High-dimensional Representation. In: Boll, S., Tian, Q., Zhang, L., Zhang, Z., Chen, Y.-P.P. (eds.) *MMM 2010. LNCS*, vol. 5916, pp. 358–369. Springer, Heidelberg (2010)
5. McDonal, C.: Machine Learning: A Survey of Current Techniques. *Artificial Intelligence Review* 3, 243–280 (1989)
6. Komorowski, J., Øhrn, A., Skowron, A.: The ROSETTA Rough Set Software System. In: Klösgen, W., Zytkow, J. (eds.) *Handbook of Data Mining and Knowledge Discovery*, ch. D.2.3. Oxford University Press, Oxford (2002)
7. Vapnik, V.: *The Nature of Statistical Learning Theory*, 2nd edn. Springer, Heidelberg (1999)
8. Hsu, C., Chang, C., Lin, C.: A Practical Guide to Support Vector Classification, <http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>
9. Breiman, L.: Bagging Predictors. *Machine Learning* 24(2), 123–140 (1996)
10. Ho, T.: The Random Subspace Method for Constructing Decision Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(8), 832–844 (1998)
11. Sande, K., Gevers, T., Snoek, C.: Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(9), 1582–1596 (2010)
12. Smeaton, A., Over, P., Kraaij, W.: Evaluation campaigns and TRECVID. In: Proc. of MIR 2006, pp. 321–330 (2006)
13. Tao, D., Tang, X., Li, X., Wu, X.: Asymmetric Bagging and Random Subspace for Support Vector Machines-based Relevance Feedback in Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(7), 1088–1099 (2006)
14. Saha, S., Murthy, C., Pal, S.: Rough Set Based Ensemble Classifier for Web Page Classification. *Fundamenta Informaticae* 76(1-2), 171–187 (2007)
15. Guo, G., Dyer, C.: Learning from Examples in the Small Sample Case: Face Expression Recognition. *IEEE Transactions on Systems, Man and Cybernetics - Part B* 35(3), 477–488 (2005)

Community Discovery from Movie and Its Application to Poster Generation*

Yan Wang^{1,2}, Tao Mei¹, and Xian-Sheng Hua¹

¹ Microsoft Research Asia, Beijing 100190, P.R. China

² University of Science and Technology of China, Hefei 230027, P.R. China
`grapeot@mail.ustc.edu.cn, {tmei,xshua}@microsoft.com`

Abstract. Discovering roles and their relationship is critical in movie content analysis. However, most conventional approaches ignore the correlations among roles or require rich metadata such as casts and scripts, which makes them not practical when little metadata is available, especially in the scenarios of IPTV and VOD systems. To solve this problem, we propose a new method to discover key roles and their relationship by treating a movie as a small community. We first segment a movie into a hierarchical structure (including scene, shot, and key-frame), and perform face detection and grouping on the detected key-frames. Based on such information, we then create a community by exploiting the key roles and their correlations in this movie. The discovered community provides a wide variety of applications. In particular, we present in this paper the automatic generation of video poster (with four different visualizations) based on the community, as well as preliminary experimental results.

Keywords: Content-based movie analysis, social network, video poster.

1 Introduction

With the blooming of movie industry, huge amount of movies as well as sitcoms are produced every day. It becomes important to index and search the movie library. In a movie, roles are naturally the center of audiences' interests, and the interactions among the roles help to narrate a story. Therefore, discovering the community, i.e., identifying key roles and discovering their relationship, is critical for content-based movie analysis, which can in turn benefit various applications such as summarization and browsing.

However, discovering a community in a movie is very challenging. First, character identification, which is usually the first step for relationship discovering, is difficult due to the complex environment in movies, especially the variation of characters' poses, illumination conditions, and so on [1]. Second, the correlations among different roles are hard to analyze thoroughly. This is because roles can interact in different ways, including *direct* interactions such as the dialogs with each other, and *indirect* interactions such as talking about other roles. A lot of research efforts have been focused on this topic. For identifying roles, Satoh *et al.* presented

* This work was performed at Microsoft Research Asia.



Fig. 1. An example of community graph and the generated posters from a movie. Each node in (a) denotes a key role, while edge denotes the relationship between two roles. We only show two poster styles in (b) and (c).

the first attempt on associating names and the corresponding roles in news videos based on their co-occurrence [2]. Everingham *et al.* [3] proposed a method for taking advantage of various features, including face appearance, clothes appearance, speaking status, and scripts. In addition to purely based on video contents, Liu *et al.* leveraged the image search results from Google as an extra information source to identify characters in movies [4]. In [5], Zhang *et al.* built affinity networks on faces and names independently and then match these two networks so as to assign a name to each face. RoleNet analyzes the correlations among roles by employing social network analysis, and presents two applications on leading roles identification and community discovery [6].

However, most existing research requires rich metadata, such as casts, scripts, or the crowdsourcing knowledge from the web, which is not always available in practice. This is particularly true in the interactive IPTV applications like AT&T U-verse TV [7] and a VOD system, where we can only get a brief title of each video section, but need to provide descriptions about the video. Therefore, we focus in this paper the problem of discovering key roles and their correlations when only video stream is available.

To solve the problem, we first detect the hierarchical structure (i.e., scene, shot, and key-frame) of the video, followed by detecting and grouping the faces appearing in the key-frames, and then construct a community graph based on their co-occurrence in the movie. In the graph, key roles work as vertices and their relationship works as edges. The discovered community leads to a wide variety of applications. In particular, we focus in this paper the automatic generation of video poster, which is a kind of visualization of movie content in a static preview. To generate a poster, we first detect key roles from the discovered community, select representative images for each role and typical background of the movie, and then propose four different visualization techniques based on the representative key roles and background. Fig. 1 shows a sample community extracted from a movie and generated posters. Our contributions are twofold: (1) we propose an approach to discover key roles and their relationship without rich metadata, which could be useful in many practical scenarios like IPTV and VOD systems, and (2) we present the visualization application based on the discovered community—automatic generation of movie posters in different styles.

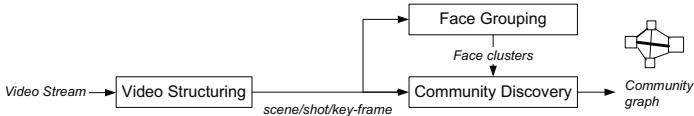


Fig. 2. Approach overview for discovering community from a movie

The remaining of the paper is organized as follows. Section 2 introduces the proposed approach. Section 3 provides the application of movie poster. Section 4 presents the experimental results, followed by the conclusions in Section 5.

2 Approach to Community Discovery from a Movie

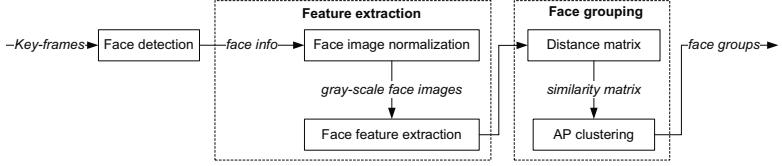
Fig. 2 shows the three major steps to discover the community from a movie. The first step is video structure analysis, which segments a video stream into hierarchical levels, including scene, shot, and key-frame. The second step detects faces from the key-frames and performs face grouping. We normalize the detected faces into (64×64) gray scale images, concatenate pixel values of each image as a feature vector, and then employ Affinity Propagation (AP) to cluster face images, with each cluster indicating a role in the movie. Based on these roles and video structure information (i.e., the co-occurrence of the roles in a scene), the third step constructs a community graph, in which each vertex indicating a key role and the weight of edges indicating relationship between each pair of roles. Specifically, we take the co-occurrence of the roles in a scene as the signal of their correlations.

2.1 Face Grouping

Before we perform face grouping, a preliminary step is employed to segment a movie into an hierarchical structure, from large to small, scene, shot, and key-frame¹ [8]. Then, we used the multi-view face detector to detect face region in each key-frame [9]. The approach to grouping faces is shown in Fig. 3. First, we detect faces from the key-frames to get the bounding face rectangles as face images. Then, the face images are normalized into 64×64 gray-scale images. For each face image, the gray value of all pixels are concatenated as a 4096-dimensional vector. We also investigate the learning-based face descriptor normalized by PCA (LE-PCA) as feature vector [9]. We then employ Affinity Propagation (AP) to cluster these images [10].

AP is a clustering algorithm taking a similarity matrix as input and outputting clusters with an exemplar for each cluster. It fits our situation in that, unlike other clustering algorithms such as K-Means, AP does not need the pre-defined number of clusters. In addition, it can provide an exemplar for each cluster, which benefits the visualization of the community and further applications. The similarity of each two vectors are calculated using their Euclidean

¹ Each shot only has one key-frame in this work.

**Fig. 3.** Approach to face grouping

distance. Suppose we have n face images $\mathcal{F} = \{f_i\}_{i=1}^n$, AP initially treats each face as a potential exemplar of itself. Then, it uses an iterative way to calculate the exemplar for each cluster. In brief, for each pair f_i and f_j , two kinds of information are propagated. One is $r(i, j)$ called “responsibility,” transmitted from f_i to f_j and indicating how well f_j would act as an exemplar of f_i among all the potential exemplars of f_i . The other is $a(i, j)$ called “availability,” transmitted from f_j to f_i and indicating how appropriate f_j would act as an exemplar of f_i by considering other potential exemplars which may choose f_j as an exemplar. Given the similarity matrix $S_{n \times n} = \{s_{i,j} | s_{i,j}$ is similarity between f_i and $f_j\}$, these two kinds of messages are propagated iteratively as,

$$\begin{aligned} r(i, j) &\leftarrow S_{i,j} - \max_{j \neq j'} \{a(i, j') + s_{i,j'}\} \\ a(i, j) &\leftarrow \min \{0, r(j, j)\} + \sum_{i' \notin \{i, j\}} \max \{0, r(i', j)\}. \end{aligned} \quad (1)$$

The self availability is determined by

$$a(j, j) \leftarrow \sum_{i' \neq j} \max \{0, r(i', j)\}. \quad (2)$$

The iteration process will stop when convergence is reached. Then the exemplar of each face f_i is extracted by solving

$$\arg \max_j \{r(i, j) + a(i, j)\}. \quad (3)$$

Faces with the same exemplars are clustered as a group. Each group is regarded to contain the faces of one role. Fig. 4 shows an example of face grouping results from a comedy sitcom.

2.2 Community Graph Construction

To describe the interactions among the roles in a movie, we adopt social network analysis—a research field in sociology which models interactions among people as a complex network among entities and tries to discover inside properties [11]. People in the society or roles in a movie are represented by vertices in a social network, while the correlations among people are modeled as weighted edges. One of the critical issues is how to model the relationship among different roles. In a movie, roles interact in various ways, such as body contacts, talking



Fig. 4. A face grouping result from a comedy sitcom. Each row corresponds to a group.

with each other, and appearing in dialogs. Considering movie is an art using shots/scenes to narrate the story, as well as we are not given any metadata (e.g., scripts), we choose “visually accompanying” as the sign for roles to interact. Specifically, if two roles appear in the same scene, we will regard that they have correlation. Furthermore, the closer the roles appear in the time line, the stronger relationship they have. According to above analysis, we treat the correlations $d(a, b)$ between two faces a and b as

$$d(a, b) = \begin{cases} c/(1 + \Delta T) & \text{when face } a \text{ and face } b \text{ are in the same scene} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

in which c is a constant in seconds, and $\Delta T = |time(a) - time(b)|$ measures the temporal distance of two faces.

By collecting the correlations of all the faces from each role, we can get the weight of the edge between role (cluster) A and B in the graph, and thus get the adjacent matrix $W_{A,B}$ by

$$W_{A,B} = w(A, B) = \sum_{a \in A} \sum_{b \in B} d(a, b) \quad (5)$$

Typically, we can detect around 500 faces from the keyframes of a two-hour movie, which means we need to calculate $d(a, b)$ for about $C_{500}^2 \approx 10^5$ times. Considering that each pair of faces a and b from different scenes should have $d(a, b) = 0$, a lot of computations can be neglected. Therefore, we calculate $d(\cdot)$ scene by scene, i.e., only the faces in the same scene are included in the computation. A sample community graph is shown in Fig. 11. Note that the size of each vertex stands for the size of the corresponding face group. The width of each edge indicates its weight. Only the edges with enough weights and the vertices connected by these edges are rendered.

3 Application to Video Poster Generation

The discovered community can provide a wide variety of applications, e.g., summarization, visualization, and so on. In particular, this paper is concerned with



Fig. 5. Examples of professional movie posters

the application of video summarization. We present video poster technique which is able to automatically create four different types of posters based on the community graph. Note that video poster is defined as a static preview (either an existing image or a synthesized image) of movie content.

In the domain of movie and TV series, a poster is an image (often containing texts) designed to promote the movie. Its task is to attract audience's attention as well as reveal important information about the movie. According to [12], a good movie poster is characterized by: (1) having conspicuous main theme and object; (2) attracting audience's attention by shocking colors and texture; (3) being self-contained and self-explained; (4) being specially designed for viewed from a distance, especially on narrative texts and figures. Fig. 5 shows some examples of professional movie posters.

Based on these design principles, we provide posters with four different styles, i.e., (1) one representative frame, (2) Video Collage style [13] [14], (3) Picture Collage style [15], and (4) synthesized style. Since roles are the center of movies and often make up the main part of posters, we will first identify the key-roles, and then generate these four types of posters.

3.1 Key Role Identification

According to above analysis, a poster should contain key roles, which can be selected from the community graph obtained in Section 2. Usually, the key roles: (1) appear frequently in the movie, and (2) have many interactions with other roles. The problem of key role identification is then transferred to find vertices that contain the most popular faces and with good connections to other vertices. Thus, we define an “role importance” function $f(v)$ on a vertex v :

$$f(v) = \text{FaceNum}(v) + \lambda \bar{\lambda} \text{Degree}(v), \quad (6)$$

where $\text{FaceNum}(v)$ denotes the number of faces in the cluster v , and $\text{Degree}(v)$ is the degree of the vertex v in the graph (i.e., sum of the weight of the connected edges with v). Since $\text{FaceNum}(v)$ and $\text{Degree}(v)$ are usually in the different granularity, $\bar{\lambda} = \frac{\text{num of faces}}{\sum_v \text{Degree}(v)}$ is introduced to balance these two terms. λ is a constant to adjust the importance of the two terms. Considering the appearance of posters, we choose 3-5 roles with the largest $f(v)$ as the key roles.

3.2 Poster Generation

A Representative Frame as Poster. One straightforward method is to provide a representative frame as the poster. A frame from the movie can preserve the real scene and roles. To enable the poster to reveal the main theme and subjects so as to satisfy the poster design principles, we add the following constraints on the selected frame: (1) it should contain as many major roles as possible; (2) it should fit the whole movie in color theme; (3) it should have good visual quality. We then define the function $r(f_i)$ on the key-frame f_i and select the frame with the maximum r :

$$r(f_i) = \sum_j \frac{\log S(f_i^{(j)})}{|h(f_i) - \bar{h}|} \quad (7)$$

where j indicates the face index in the frame f_i , $S(f_i^{(j)})$ denotes the area of j -th face, $h(f_i)$ indicates the color histogram of key-frame f_i , and \bar{h} is the average one. Other features related to video quality could be also integrated [8].

A Collage as Poster. Collage is a compact and visually appealing way to organize the key roles [13] [14] [15]. With the enlarged face area of roles, collages often have conspicuous themes and can attract audiences easily. To generate a collage-style poster, we first extract a representative face for each key role, and then use the collage techniques to organize the faces into a visually appealing poster. Note that in this category, we mention two kinds of collages—Picture Collage [15] or Video Collage [13] [14]. The representative faces are expected to be frontal, visible and clear enough. Therefore, we select the face with largest area among all the frontal ones. We crop the area containing the face as the material to propose the collage. For Picture Collage style poster, we set the face region as the region-of-interest (ROI) and employ the Markov Chain Monte Carlo (MCMC) to search for a collage, in which all ROIs are visible while other parts are overlaid. For Video Collage style posters, we concatenate the images as a big poster, and smooth the boundaries to make it more natural.

Synthesized Poster. A more professional style is synthesized posters. As shown in Fig. 5, a synthesized poster is usually generated by seamlessly embedding the images of key roles on a representative background. It contains not only representative background which can help introduce the typical scenes and story background, but also featured roles to attract audiences [16]. To create such a poster, we first extract a representative background key-frame, filter out objects from the background with/without user interactions, and then seamlessly merge the portrait images of key roles into the background.

The selection of background key-frame is similar to the process of selecting a representative frame as poster. The difference is that fewer faces are expected to appear in the background so that audience's attention would not be distracted. Therefore, we select the frame with the minimum $r(f_i)$ as background, defined in equation (7). From the background image, the foreground objects are extracted with/without user interactions to distract user attention toward the background.

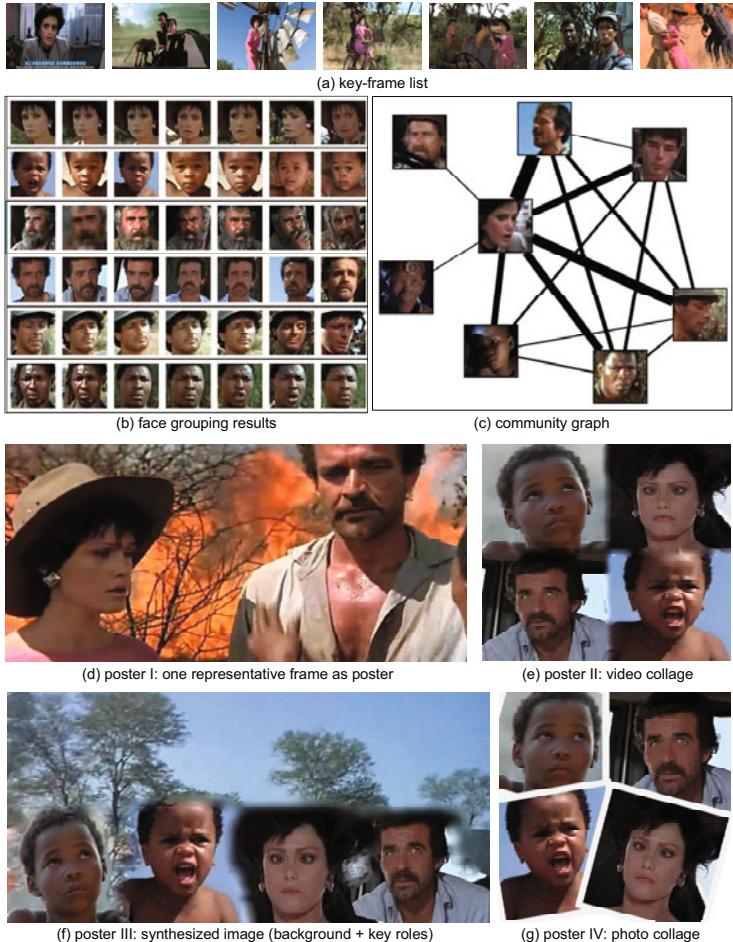


Fig. 6. An example of posters from the movie “The gods must be crazy”

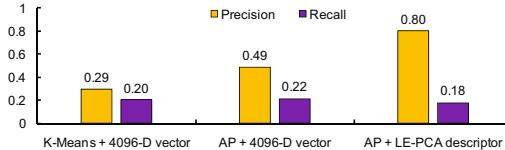
We employ the intelligent background removal for the implementation of foreground removal [17]. Once we obtain the background image (without foreground objects), we then seamlessly merge the normalized portraits or the key roles horizontally on it. The position and scale of portraits are decided by the size of the corresponding face group in the community graph. Figure 6 shows the results of posters from a movie.

4 Experiment

We collected two Chinese sitcoms recorded from CCTV-1 (China Central Television Station—Channel 1), one English sitcom, and six English movies. The movies and sitcoms cover diverse genres including kongfu, romantic, comedy, and science fictions. The detailed information is shown in Table 1. We first evaluate face

Table 1. Information about experimental data

ID	Type	Duration	Genre	Language	# of Faces	# of Clusters
1	Sitcom	46m	Kungfu	Chinese	214	9
2	Sitcom	48m	Romantic	Chinese	394	14
3	Sitcom	21m	Comedy	English	304	9
4	Movie	2h23m	Adventure, Drama	English	714	11
5	Movie	2h20m	Crime, Thriller	English	487	19
6	Movie	1h57m	Drama, Comedy	English	705	27
7	Movie	1h37m	Action, Comedy	English	541	12
8	Movie	1h58m	Crime, Thriller	English	457	15
9	Movie	1h33m	Sci-Fi, Action	English	810	30

**Fig. 7.** Comparison on the performance on face grouping from different algorithms

grouping and key role detection results in terms of precision and recall, and then invite several subjects to do a user study on the quality of the posters.

4.1 Evaluation of Face Grouping

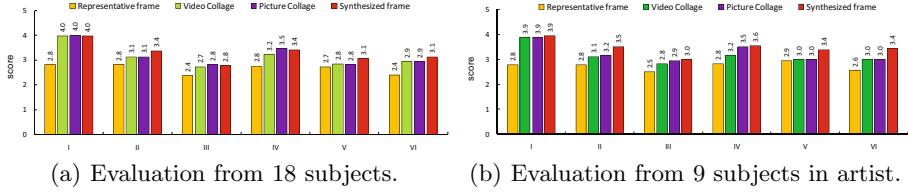
We invited two labelers majored in artistic to group the faces in the movies manually. They were required to first watch the whole movie/sitcom thoroughly, and then classify all faces into groups. The number of detected faces and manually grouped clusters is shown in Table. 1. Since the number of face groups recognized by our proposed approach may be different from that from the annotated ground truth, we first selected five largest groups from both the ground truth and our results, and then matched them with a greedy algorithm. Then, for each group G from the ground truth and group R from our result, the precision and recall are calculated by

$$\text{Precision} = |G \cap R| / |R|, \quad \text{Recall} = |G \cap R| / |G|. \quad (8)$$

Then, we average the precision and recall of the five groups as the evaluation for each movie. For the nine movies we selected, the average precision for face grouping is 0.80, while the recall is 0.18. Note that in our problem, precision is much more important than recall, since only the recognized key roles (usually three, four, or five roles) will be used for poster generation. We also tried other clustering algorithms, i.e., K-means and AP [10], with the different features. Fig. 7 shows the comparison. We can see that “AP + LE-PCA descriptor” has much better performance in terms of precision, while recalls of the three algorithms are comparable.

4.2 Evaluation of Key Role Extraction

Based on the annotated face grouping results, the two labelers then selected some key roles from the movies. The key roles are defined as major characters

**Fig. 8.** Results of user study towards poster quality

that lead the plot of the movie. The numbers of key roles are determined by how they understood the movies/sitcoms. The selected roles act as the ground truth. With the standard key-roles G as the ground truth and the extracted key-roles R , we calculate the precision and recall for each movie by equation (8). The average precision is 0.75 while the recall is 0.87.

4.3 Evaluation of Poster Generation

To evaluate the effectiveness of the generated posters, we conducted a subjective user study. We invited 18 subjects, nine of which are with artistic background (including three males and six females). First, they were asked to watch the movie/sitcom, and then provided with the posters in four different styles. At last, they were asked to answer the following six questions about the posters with a score from 1 to 5, while 1 is “absolutely no” and 5 “absolutely yes.” The six questions are:

- (I) Do you think the poster indicates the actual major roles of the movie?
- (II) Do you think the poster compatible with the movie in genre?
- (III) Do you think the poster can represent the main content of the movie?
- (IV) Do you think the layout of the poster comfortable?
- (V) Do you think the poster can attract you easily?
- (VI) Do you think the poster good in an overall manner?

Fig. 8 shows the evaluation results. We can see the average satisfaction scores increase from the simplest representative frame style, to collage style, and then to the more professional synthesized style. Except representative frame style, the average score to question (I) about key role identification is high (i.e., near 4), which has proved the effectiveness of community extraction. The reason that the representative frame style cannot achieve a higher score in this question is that there does not always exist a single frame containing all the key roles in the movie. However, the other poster styles solve this problem by sacrificing some information about typical scenes. We can also see the the scores from the nine subjects majored in artist are around 0.3 higher than those on average.

5 Conclusion

In this paper, we propose a new perspective for movie content analysis without rich available metadata, where conventional methods cannot work well. We first

analyze video structure, then group faces, and finally extract a community to reveal the correlations among different roles. Based on the community graph, we demonstrate a new application to automatically generate a collection of posters for a movie. Future work may include more investigating temporal information in the community extraction to improve the poster and automatic speech recognition for key role detection.

References

1. Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.J.: Face recognition: A literature survey. *ACM Computing Surveys* 35(47), 399–458 (2003)
2. Satoh, S., Kanade, T.: Name-It: Association of face and name in video. In: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition, San Juan, Puerto Rico, pp. 368–373 (1997)
3. Everingham, M., Sivic, J., Zisserman, A.: Hello! My name is Buffy—automatic naming of characters in TV video. In: Proceedings of the British Machine Vision Conference (2006)
4. Liu, C., Jiang, S., Huang, Q.: Naming faces in broadcast news video by image google. In: Proceeding of ACM International Conference on Multimedia, Vancouver, Canada, pp. 717–720 (2008)
5. Zhang, Y.F., Xu, C., Lu, H., Huang, Y.M.: Character identification in feature-length films using global face-name matching. *Trans. on Multimedia* 11(7), 1276–1288 (2009)
6. Weng, C.Y., Chu, W.T., Wu, J.L.: RoleNet: treat a movie as a small society. In: Proceedings of the international Workshop on Multimedia Information Retrieval, Augsburg, Bavaria, Germany, pp. 51–60 (2007)
7. AT&T: U-verse tv, <http://www.att.com/u-verse/>
8. Mei, T., Hua, X.S., Zhu, C.Z., Zhou, H.Q., Li, S.: Home video visual quality assessment with spatiotemporal factors. *IEEE Trans. on Circuits and Systems for Video Techn.* 17(6), 699–706 (2007)
9. Cao, Z., Yin, Q., Tang, X., Sun, J.: Face recognition with learning-based descriptor. In: Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (2010)
10. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. *Science* 315, 972–976 (2007)
11. Scott, J.P.: Social Network Analysis: A Handbook. SAGE Publications (2000)
12. Frascara, J.: Communication design: principles, methods, and practice. Allworth Communications, Inc. (2004)
13. Mei, T., Yang, B., Yang, S.Q., Hua, X.S.: Video collage: Presenting a video sequence using a single image. *The Visual Computer* 25(1), 39–51 (2009)
14. Wang, Y., Mei, T., Wang, J., Hua, X.S.: Dynamic video collage. In: International Conference on MultiMedia Modeling, Chongqing, China, pp. 793–795 (2010)
15. Wang, J., Sun, J., Quan, L., Tang, X., Shum, H.Y.: Picture collage. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 347–354 (2006)
16. Skolos, N., Wedell, T.: Type, Image, Message: A Graphic Design Layout Workshop. Rockport Publishers (2006)
17. Liu, T., Sun, J., Zheng, N.N., Tang, X., Shum, H.Y.: Learning to detect a salient object. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)

A BOVW Based Query Generative Model

Reede Ren¹, John Collomosse¹, and Joemon Jose²

¹ CVSSP, University of Surrey, Guildford, GU2 7XH, UK

² IR Group, University of Glasgow, Glasgow, G12 8QQ, UK

Abstract. Bag-of-visual words (BOVW) is a local feature based framework for content-based image and video retrieval. Its performance relies on the discriminative power of visual vocabulary, *i.e.* the cluster set on local features. However, the optimisation of visual vocabulary is of a high complexity in a large collection. This paper aims to relax such a dependence by adapting the query generative model to BOVW based retrieval. Local features are directly projected onto latent content topics to create effective visual queries; visual word distributions are learnt around local features to estimate the contribution of a visual word to a query topic; the relevance is justified by considering concept distributions on visual words as well as on local features. Massive experiments are carried out the TRECvid 2009 collection. The notable improvement on retrieval performance shows that this probabilistic framework alleviates the problem of visual ambiguity and is able to afford visual vocabulary with relatively low discriminative power.

1 Introduction

Bag-of-visual words (BOVW) is a promising framework for content-based image and video retrieval. Key points are collected from images to denote salient regions; local features such as SIFT [8] are computed around key points and are clustered to generate a visual vocabulary; each SIFT cluster is registered as a unique visual word; an image is therefore translated into a set of visual words; appearance statistics, *e.g.* the histogram of visual word frequency [5], is calculated to decide on the contribution of visual words as well as to estimate the relevance. Fei-Fei *et al.* [6] value BOVW as a reliable approach to bridge the gap between low-level visual characters and high-level image contents. They demonstrate the effectiveness in general scene discrimination and visual object modelling. The recent TRECvid competition also reports that BOVW achieves the best performance in the automatic search task, although the overall precision remains unsatisfied [12][13].

One of the major challenges in BOVW is the generation of visual vocabulary. Since visual words do not naturally exist in images, vocabulary generation is an optimisation across the local feature set of the document collection, which tries to maximise the discriminative power while not reduces retrieval efficiency. Given the extremely huge number of local features and the sparsely distributed visual contents, such an optimisation process is of a high complexity and is not so robust

[13]. K-mean clustering and its variants are widely adopted as a low cost solution but requires an extra preliminary of vocabulary size, *i.e.* cluster number [16]. Jiang *et al.* [5] carried out massive experiments on the TRECVID 2006 collection and evaluated various vocabulary size from 200 to 10,000. The authors observe the problem of visual ambiguity that a local feature can be assigned to multiple visual words and report no vocabulary configuration out-performs the others. This indicates that visual ambiguity weakens the effectiveness of visual words for content discrimination and that the visual vocabulary generated by K-mean clustering is unreliable for retrieval [12]. Moreover, appearance statistics is closely associated with the vocabulary. In addition, Li *et al.* [6] point out that a high-level visual concept is contributed by a set of rather than a single local feature. An improperly configured vocabulary not only misguides appearance statistics but also leads to the problem of ‘double counting’ in pattern recognition, as a visual concept is randomly projected onto multiple visual words.

This paper aims to relax the constraints from visual vocabulary and thus to alleviate related problems such as visual ambiguity and visual word assignment. We propose the probabilistic framework of query generative model for BOVW based retrieval. Latent query topics are learnt directly from local features. This avoids the usage of visual vocabulary in query generation and results in a precise query modelling. Moreover, local feature based query models allows the estimation of visual word contributions to a local feature. A new soft assignment scheme is therefore developed by exploiting neighbourhood statistics of visual words around a local feature. Experimental results show that this new assignment scheme is more effective than cluster centre based assignment. In addition, the query generative model make possible the introduction of prior/external knowledge, *i.e.* local feature based visual concept model, and promises a further improvement on retrieval performance. Some research issues are also addressed, *e.g.* shot-based temporal accumulation.

The remainder of this paper is organised as follows. Section 2 briefly reviews components in the framework of BOVW, including key point detection, visual vocabulary generation and relevance estimation. Section 3 explains the query generative model and describes the estimation of (1) local feature distribution, (2) latent topic distribution among query examples; and (3) local feature distribution in a latent topic. The optimised number of query topics is learnt by maximising the entropy of latent topic distribution. Experimental results are stated in Section 4, including experimental configuration, baseline creation and shot-based retrieval performance on the TRECVID 2009 collection. Conclusions are found in Section 5.

2 Related Work

In this section, we briefly review the development of BOVW in content-based video retrieval. BOVW was originally proposed for visual object recognition [4]. Snoek *et al.* [12] introduce this framework into content-based video retrieval and

treat detection confidence as an effective relevance estimator. The authors claim that BOVW is a possible gateway to the robust semantic video retrieval. Three key issues are then identified in [5], including key point detection, vocabulary generation and visual word weighting scheme.

Key points are essential small image regions for the recognition of visual contents. Detection algorithms such as Harris Laplace, Boost Colour Harris Laplace, Difference of Gaussian (DOG) and grid sampling are compared in [13]. Uijlings *et al.* [13] justify their choice by retrieval performance on the TRECVID 2006 collection. The authors recommend grid sampling for discriminative power despite the largest key point collection among all. We notice that the performance difference is insignificant to support such a computational cost. Therefore, we use the DoG key point detector [8] in this paper to balance the effectiveness and the efficiency. Later, Battiatto *et al.* [2] group key points to form visual synset, the counterpart of n-gram representation in textual documents. The authors show that visual synset is more invariant and more discriminative than a single key point in object categorisation. However, the collection of possible visual synsets increases exponentially with the average synset length. Liu *et al.* [7] explore Ababoost to select the most discriminative combinations. Savarese *et al.* [11] compute correlation matrixes to identify the most common key point co-occurrence. Zhang *et al.* [16] apply geometric constraints to group key points nearby. These works decreases the blooming speed to sub-exponent.

Vocabulary generation is usually regarded as a problem of unsupervised learning, which is closely associated with two factors: (1) the distribution of visual concepts and (2) the distribution of local features (Figure 1b). The number of visual concepts decides on the necessary size of a visual vocabulary, although this requirement is latent. This is due to the complexity in the estimation of the actual concept number in general purpose retrieval. Local feature distribution defines the boundary between visual words and therefore affects the discriminative power of vocabulary. Marsezlek *et al.* [9] propose K-mean clustering with a random cluster number for efficiency. Uijling *et al.* [13] adapt the random forest as a robust replacement for K-mean. Some dimensionality removal approaches are also tried to improve the effectiveness, *e.g.* principle component analysis [13]. However, it remains a research question how to optimise a visual vocabulary, given the huge feature collection.

BOVW adapts textual term weighting schemes for relevance estimation, *e.g.* term frequency and inverse document frequency. Evering *et al.* develop a binary weighting for image based visual word matching. Marsezlek *et al.* [9] adapt a BM25 like retrieval model to estimate the relevance between video shots. Agarwal *et al.* [1] notice visual ambiguity and take the difference between local features into consideration. The authors propose a language model like probabilistic framework to simulate local feature distribution as well as to optimise the vocabulary by using the posterior mixture-component membership. Jiang *et al.* [5] propose soft assignment and label a local feature with multiple visual words at different weights. This results in a significant precision improvement on

the TRECVID 2006 collection. Germert *et al.* [14] furthermore propose a query dependent soft assignment in which the weight of a local feature to a visual word is learnt from query classes.

3 Query Generative Model

In this section, we justify our query generative framework for BOVW based retrieval. BOVW can be regarded as a four-tier document generative model (Figure IIb). A visual word denotes a distribution of local features; a visual concept is represented by a probabilistic distribution across visual words [6]; the relevance is estimated by comparing the mixed distribution of visual concepts [12]. In addition, this process can be simplified into three layers by projecting local features directly onto visual concepts, *e.g.* the visual concept ontology called ImageNet [3]. As a consequence, the relevance estimation in BOVW is to generate a visual document by picking a distribution over visual concepts and then a concept dependent distribution of visual words/local features (Figure IIa).

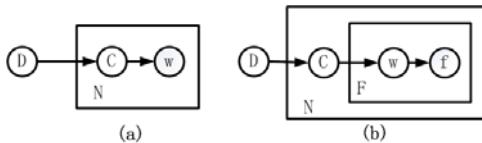


Fig. 1. BOVW based generative model, where **D** denotes visual documents, **C** stands for visual concepts, **w** and **f** are visual words and local features, respectively

Document generative models such as Latent Dirichlet Allocation (LDA) are of high complexity. Given the huge number of local features, it is computationally unfeasible to exploit document generative models. We hence turn to an alternative called query generative model (Figure 2). Zhai *et al.* [15] prove that the document generative model is theoretically equivalent to the query generative approach. Moreover, there are three advantages in the query generative model. Firstly, effective query concepts are densely distributed in query example collection. This facilitates the learning of latent visual concepts. Secondly, query example collection is small. It is unnecessary to define an individual vocabulary to conceal local feature variations. This alleviates the problem of vocabulary generation and indicates that the query generative model can afford visual vocabulary of relatively low discriminative power. Finally, the query generative model provides a precise modelling for relevance and allows the usage of prior knowledge, such as local feature based concept models in ImageNet. This will improve the effectiveness and the robustness of BOVW based retrieval. In summary, the query generative model provides a dynamic probability framework for BOVW based relevance estimation.

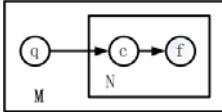


Fig. 2. Query generation model, where q denotes a query example, C and f are visual concepts and local features, respectively

The relevance estimation function (RSV) is as follows.

$$\begin{aligned} p(d|Q) &= \sum_{C_Q} p(d|C_Q)p(C_Q|Q) \\ &= \sum_{C_Q} (\sum_{f_Q} p(d|f_Q)p(f_Q|C_Q))p(C_Q|Q) \end{aligned}$$

where C_Q denotes latent concepts in a query Q ; d stands for a video document; f_Q and f_d are local feature collection of Q and d , respectively. This function shows that checking the most relevant query concepts is an effective alternative to studying all possible concepts in documents. If local features are quantified into visual words, the RSV can be rewritten into the joint set description.

$$p(d|Q) = \sum_{C_Q} (\sum_{f=f_Q \cap f_d} p(d|f)p(f|C_Q))p(C_Q|Q) \quad (1)$$

where f_d is the local feature collection in a visual document. This description is equivalent to the binary/tf-idf based weighting scheme, where $p(d|f_Q)$ stands for the binary matching between visual words. Consider visual ambiguity in the BOVW framework, f_d can be treated as a noisy derivation from f_Q . We hence improve the RSV as follows.

$$p(d|Q) = \sum_{C_Q} (\sum_{f_Q} (\sum_{f_d} p(d|f_d)p(f_d|f_Q))p(f_Q|C_Q))p(C_Q|Q) \quad (2)$$

$p(f_d|f_Q)$ is a similarity measurement, which can be an assignment scheme to highlight the contribution of local feature f_Q to visual word f_d . For example, the soft-assignment in █ is an Euclidean distance based contribution estimation, where $p(f_d|f_Q) = \frac{1}{2L_2(f_d, f_Q)}$.

To summarise, three components are to learn in the query generative approach, the concept distribution in query examples $p(C_Q|Q)$, the local feature distribution for a concept $p(f_Q|C_Q)$ and the local feature distribution in documents $p(d|f_d)$. This is because $p(d|f_d) = \frac{p(f_d|d)p(d)}{p(f_d)} \sim p(f_d|d)$, as d is uniformly distributed and $p(f_d)$ is constant. In addition, the RSV (Equation 2) indicates that the query generative approach will be more effective in shot-based video retrieval. This is because $p(f_d|d)$ would be a Boolean if only one keyframe were considered.

3.1 $p(f_d|f_Q)$

$p(f_d|f_Q)$ is a similarity measurement between local features in visual documents and in query examples. The approximation of cosine distance is ineffective because local feature descriptors, such as SIFT [8], are a combination of edge and texture histogram. Note that $p(f_d|f_Q)$ could be regarded as a soft assignment scheme. We consider the ad-hoc density of local feature distribution (Figure 3). This alleviates the preliminary requirement on vocabulary size.

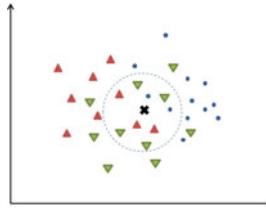


Fig. 3. Neighbourhood based soft assignment, where x denotes f_Q

We project f_Q directly into the local feature space and collect a given size of neighbourhood, e.g. 0.001% of local feature space. The appearance number of a visual word in the neighbourhood is counted and the appearance ratio is taken as the soft assignment weight.

$$p(f_d|f_Q) = \frac{\#_R(f_d)}{\#_{f_d \in R}(f_d)} \quad (3)$$

where R is the neighbourhood of f_Q . The usage of pyramid index ensure this computation of a constant complexity.

3.2 $p(f_d|d)$

$p(f_d|d)$ counts the appearances of local features in a video document d . Since not all SIFTS could be matched frame by frame, we develop a two-pass strategy to improve the robustness, including SIFT matching and block-tracking. SIFT matching compares 128-dimensional SIFT values to couple local features. If not matched, we turn to the step of block tracking. A visual frame is re-sized by affine parameters in SIFT descriptor. A $4 * 4$ block is collected around the point and searched in nearby areas of sequent frames. If the colour-based block distance is small enough, we think this local feature remains though ignored by the key point detector. Hence, $tf = tf_{sift} + atf_{tracking}$, where tf_{sift} and $tf_{tracking}$ are matching count by sift matching and block tracking, respectively; a is 0.75 learnt by experiments. $p(f_d|d)$ is estimated as follows.

$$p(f_d|d) = \frac{tf}{\|d\| + s} + \epsilon \quad (4)$$

where $\|d\|$ is the number of key points in a shot, s denotes the size of low-level feature terms [10] and ϵ is a smoothing parameter. Two types of feature terms, colour layout and edge histogram, are used in experiments to describe global characters.

3.3 $p(C_Q|Q)$ and $p(f_Q|C_Q)$

$p(C_Q|Q)$ and $p(f_Q|C_Q)$ aim to fit a given number of visual concepts to query examples. In the TRECVID collection, a query consists of 7-11 images and a short textual description. We take the number of textual noun entities as the initial estimation of possible concept number. The example set is re-sampled to create a serial of constant size sub-query collections ($q_1 \dots q_M$). We compute $p(f|q_i)$ and Bootstrap is used to improve the estimation. $p(f|q_i)$ is computed by the EM algorithm.

E-Step

$$p(c_l|q_i, f_j) = \frac{p(f_j|c_l)p(c_l|q_i)}{\sum_{l=1}^K p(f_j|c_l)p(c_l|q_i)} \quad (5)$$

M-Step

$$p(f_j|c_l) = \frac{\sum_{i=1}^M t f_j(q_i) p(c_l|q_i, f_j)}{\sum_{j=1}^N \sum_{i=1}^M t f_j(q_i) p(c_k|q_i, f_j)} \quad (6)$$

$$p(c_l|q_i) = \frac{\sum_{j=1}^N t f_j(q_i) p(c_l|q_i, f_j)}{\sum_{j=1}^N t f_j(q_i)} \quad (7)$$

where N is the number of local features and K the number of relative concepts. The sum of feature entropy is calculated to decide on the number of effective concepts. In addition, the actual query concept number is less than five in the TRECVID 2009 collection.

$$\|C_Q\| = \arg \max_K \sum_{l=1}^K \sum_{j=1}^N (-p(f_j|c_l) \log p(f_j|c_l)) \quad (8)$$

4 Experiment

The TRECVID 2009 collection is used for evaluation, which consists of 219 videos and 21 queries. Every video is made up by about 150 shots and the overall shot number is 212,256. We sample a shot at 1/10. DOG is used to detect key points and 128-dimensional SIFT to describe local characters. The SIFT collection contains about 17M samples. Retrieval performance is measured by the TRECVID evaluation tool [2]. The following sections will address the configuration of retrieval system, including soft assignment scheme, shot-based relevance and visual vocabulary generation, and finally compare the performance with LDA and the best record in the competition.

4.1 Soft Assignment

Soft assignment is an efficient approach to alleviate visual ambiguity. In this paper, we exploit the neighbourhood statics around a local feature. There are two issues affecting the effectiveness: neighbourhood scope and vocabulary size. The choice of neighbourhood is a trade-off between the efficiency and statistical robustness. On one hand, a large neighbourhood ensures statistical robustness at the cost of efficiency by reaching more samples than a small one does. On the other hand, a small neighbourhood needs less efforts in indexing but requires an extra smoothing to avoid too many zero weights.

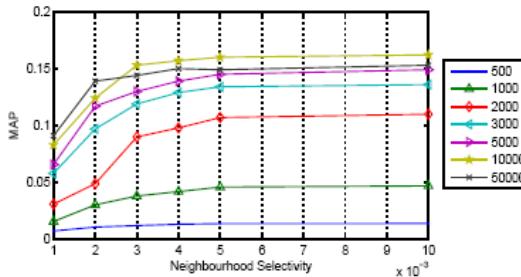


Fig. 4. MAP over neighbourhood selectivity and vocabulary size

Figure 4 compares the retrieval performance with numerous neighbourhood scope from 0.0001% to 0.01% of the SIFT collection space under all vocabulary configurations. Experimental results show that a large neighbour is able to increase precision, although the improvement will diminish after the neighbourhood is larger than 0.003%. This shows the neighbourhood based soft assignment is robust and effective. Another interesting observation is about vocabulary size. Figure 4 shows that a large vocabulary also improves retrieval performance even though a small neighbourhood is used. This consistence indicates that neighbour based assignment can be improved by a discriminative vocabulary and that the query generative model alleviates the problem of visual ambiguity. In addition, it costs about 17 milliseconds to collect samples of 0.005% feature space in pyramid index. We therefore use the neighbourhood of 0.005% to compute the soft assignment weight and the vocabulary of 10,000 in the following experiments.

Table 1 compares two soft assignment schemes, distance-based [5] and neighbourhood based. Neighbourhood based scheme is significantly better than distance-based.

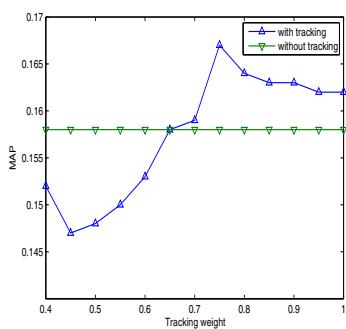
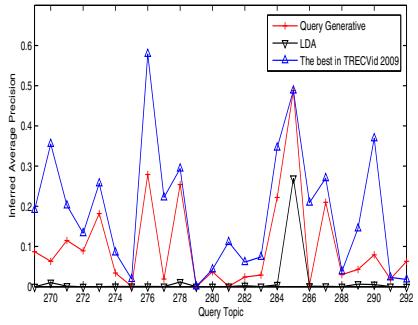
4.2 Shot-Based Relevance

An obvious difference between video and image retrieval is temporal continuity inside media documents. Many video retrieval systems regard a shot as an image collection and represent a shot by accumulating local features from key frames. Figure 5 compares performances under various weights from 0.4 to 1 on local

Table 1. MAP over soft assignment scheme

Visual word number	Distance-based @ 5 classes [5]	Neighbourhood-based @ 0.005%
500	0.008	0.014
1000	0.010	0.046
2000	0.014	0.107
3000	0.015	0.134
5000	0.019	0.145
10,000	0.021	0.167
50,000	0.012	0.149

feature tracking. A 4.38% improvement is observed over the best performance on key frame based retrieval (10,000 visual words and neighbourhood of 0.005% feature space). However, an improper weight may also decrease the performance. This indicates that the improvement from temporal tracking might be limited. There are two possible explanations: (1) an intensive key frame sampling is good enough for shot representation in retrieval and (2) query topics in the TRECVID 2009 mostly focus on a static scene rather than a continuous motion.

**Fig. 5.** MAP over shot accumulation**Fig. 6.** Top AP in the TRECVID 2009 collection

4.3 Retrieval Performance

We use the latent Dirichlet allocation (LDA) as the baseline [10]. LDA is one of the most widely used generative model and finds applications in textual segmentation, multi-class face classification and spoken letter recognition. We use the TRECVID 2008 collection to train LDA with the latent topic number 50. 10,000 visual words are assigned to each latent topic by using reserve-jump Markov Monte Carlo. Then, a query is projected onto latent topics and the relevance is estimated by maximising the post-probability that a document to the latent topic distribution. The Merry run in the MediaMill [12] is also used as baseline due to the similarity in retrieval configuration, *e.g.* BOVW representation, feature-based retrieval and learning on-the-fly query concepts. The difference

between Merry run and ours is the estimation of relevance and the scheme of soft assignment. The Merry run casts query topics onto a group of query classes and uses support vector machine to rank documents. Table 2 compares retrieval performance by LDA, the query generative model, the Merry run in MediaMill [12] and the best-over-all in the TRECVID 2009 [10]. The best-over-all collects the best performance on every topics in the TRECVID 2009 competition, which exploits all modalities including audio scripts, high-level concepts and textual tags.

Table 2. TRECVID 2009 Performance

RUN	MAP
LDA [10]	0.0132
Query Generation	0.167
Merry Run [12]	0.089
Best over all [10]	0.188

Topic based average precision is shown in Figure 6. We achieve high scores on scene related topics, *e.g.* people at a table with a computer visible and building entrance. This shows the effectiveness of the query generative scheme in the description of multiple visual concepts.

5 Conclusion

In this paper, we adapt the query generative model for BOVW based video retrieval. This robust probabilistic framework incorporates latent visual concepts and exploits visual word distribution on local features for a superior retrieval performance. Its computation cost is lower than document generative models. This makes possible the exploitation of BOVW based generative models in a large scale retrieval. Moreover, the query generative model projects local features directly onto latent visual concepts. This alleviates the problem of visual ambiguity and relaxes the requirement on the discriminative power of visual vocabulary. System robustness is therefore improved. Nevertheless, this new BOVW-based retrieval framework facilitates the introduction of prior visual concept models into relevance estimation. This extends the scope of BOVW and allows the usage of web-based external knowledge. In short, a further improvement will be seen on retrieval performance.

References

1. Agarwal, A., Triggs, B.: Multilevel image coding with hyperfeatures. International Journal of Computer Vision 78(1), 15–27 (2008)
2. Battiatto, S., Farinella, G.M., Gallo, G., Ravi, D.: Spatial hierarchy of textons distributions for scene classification. In: Huet, B., Smeaton, A., Mayer-Patel, K., Avrithis, Y. (eds.) MMM 2009. LNCS, vol. 5371, pp. 333–343. Springer, Heidelberg (2009)

3. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR 2009 (2009)
4. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding* 106(1), 59–70 (2007)
5. Jiang, Y.-G., Ngo, C.-W.: Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval. *Computer Vision and Image Understanding* 113(3), 405–414 (2009)
6. Li, L.-J., Socher, R., Fei-Fei, L.: Towards total scene understanding:classification, annotation and segmentation in an automatic framework. In: Proc. IEEE Computer Vision and Pattern Recognition, CVPR (2009)
7. Liu, D., Hua, G., Viola, P., Chen, T.: Integrated feature selection and higher order spatial feature extraction for object categorisation. In: CVPR 2008, pp. 1–8 (2008)
8. Lowe, D.: Object recognition from local scale-invariant features. In: ICCV, pp. 1150–1157 (September 1999)
9. Marszalek, M., Schmid, C., Harzallah, H., van de Weijer, J.: Learning representations for visual object class recognition. In: ICCV (2007)
10. Punitha, P., Misra, H., Ren, R., Hannah, D., Goyal, A., Villa, R., Jose, J.M.: Glasgow university at trecvid 2009. In: TRECVID (2009)
11. Savarese, S., Winn, J., Criminisi, A.: Discriminative object class models of appearance and shape by correlatons. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006, Washington, DC, USA, pp. 2033–2040. IEEE Computer Society Press, Los Alamitos (2006)
12. Snoek, C.G.M., van de Sande, K.E.A., de Rooij, O., Huurnink, B., van Gemert, J., Uijlings, J.R.R., He, J., Li, X., Everts, I., Nedovic, V., van Liempt, M., van Balen, R., de Rijke, M., Geusebroek, J.-M., Gevers, T., Worring, M., Smeulders, A.W.M., Koelma, D., Yan, F., Tahir, M.A., Mikolajczyk, K., Kittler, J.: The mediамill TRECVID 2009 semantic video search engine. In: TRECVID (2009)
13. Uijlings, J.R.R., Smeulders, A.W.M., Scha, R.J.H.: Real-time bag of words, approximately. In: CIVR 2009, Santorini, Fira, Greece, pp. 1–8. ACM, New York (2009)
14. van Gemert, J.C., Veenman, C.J., Smeulders, A.W., Geusebroek, J.-M.: Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 1271–1283 (2010)
15. Zhai, C., Lafferty, J.: A risk minimization framework for information retrieval. *Inf. Process. Manage.* 42(1), 31–55 (2006)
16. Zhang, S., Tan, Q., Hua, G., Huang, Q., Li, S.: Descriptive visual words and visual phrases for image applications. In: ACM Multimedia 2009 (2009)

Video Sequence Identification in TV Broadcasts

Klaus Schoeffmann and Laszlo Boeszoermenyi

Institute of Information Technology, Klagenfurt University,
Universitaetsstr. 65-67, 9020 Klagenfurt, Austria
`{ks, laszlo}@itec.uni-klu.ac.at`

Abstract. We present a video sequence identification approach that can reliably and quickly detect equal or similar recurrences of a given video sequence in long video streams, e.g. such as TV broadcasts. The method relies on motion-based video signatures and has low run-time requirements. For TV broadcasts it enables to easily track recurring broadcasts of a specific video sequence and to locate their position, even across different TV broadcasting channels. In an evaluation with 48 hours of video content recorded from local TV broadcasts we show that our method is highly reliable and accurate and works in a fraction of real-time.

1 Introduction

Reliable detection of duplicates or near-duplicates of video sequences is an important issue in several application domains. For example, copyright owners of specific video sequences are interested in detecting broadcasts of their content across different TV channels. In the particular case of commercials, producers may also want to know how often the commercial is broadcasted by a specific channel and at which time it is broadcasted.

Consider the example of a music TV channel that repeatedly broadcasts a specific music clip for a number of times but show different semi-transparent overlays on top of some areas of the actual content. While recurrences of music clips in broadcasts of a specific TV channel may already vary, a different broadcasting channel may use completely different overlays and a different broadcaster logo indeed. Figure 1 shows examples of slightly different content variations of one specific music clip broadcasted by two different European music channels. While in the first row both instances of the same clip are almost equal, only the broadcasters logo at top left is slightly different, the two instances of the same clip in the second row are quite different. Here, both instances use completely different overlays of different colors and different shape. In the left example of the figure the outer black borders are occluded by inserted overlays and even the color of the broadcaster logo is different to the one used in the right image.

Figure 2 shows a more extreme example of two instances of the same clip within the same broadcasting channel. In this example the overlays occlude a quite large region of the real content and the colors of both the overlays and the logo are completely different. Reliable detection of recurring video sequences across different broadcasting channels is even more challenging, since different



Fig. 1. Slightly different content variations of the same clip (two different channels)



Fig. 2. Quite different content variations of the same clip (within the same channel)

bit-rates may be used, and different encoders may be involved, which may affect colors, brightness, contrast, and quality. Moreover, long video sequences broadcasted across different TV channels may only match partially since they may contain different interruptions due to commercial breaks. The repetition may also start or stop on different time positions, for example music TV channels often cut a few frames or seconds from the beginning or the end of a music clip.

In this paper we present a video sequence identification approach that can reliably detect recurrences of equal or similar content across different broadcasting channels and also across different qualities/encodings of a video stream. We assume a scenario where the signature of the video sequence to detect is known before the search process. For example, the producer of a commercial extracts the signature from the original content and uses it as input to the search process. We evaluated our approach with 48-hours recordings of two different European music TV channels (24-hours each), for which the ground truth has been annotated manually. For every group of repeating music clips we mark a random

instance as query and evaluate whether our method reliably and accurately finds all other recurrences.

The paper is organized as follows. Section 2 gives an overview of related work in this research field. Section 3 describes the motion-based signature that is used to find recurrences. The search algorithm is described in Section 4. In Section 5 we discuss our evaluation results before the paper is concluded in Section 6.

2 Related Work

Video sequence identification is related to video copy detection, which has raised much attention in research in the last years. One of the reason is surely that with the exponential growth of videos in our world, it becomes a challenging task for a copyright owner to track redistributions of a specific video footage. In difference to content-based video retrieval (CBVR), content-based copy detection (CBCD) aims at finding other instances of a specific video sequence. However, a *copy* is not an identical or near-replicated video sequence but rather a transformed video sequence [6]. Copies may be transformed in several different ways: color, brightness, contrast, quality, resolution, logo insertion, temporal resolution, etc. These transformations make it difficult to reliably detect all copies of a video sequence while producing no false alarm at the same time. Many content-based copy detection approaches have been proposed over the last years. A comparison of some approaches can be found in [56].

However, in difference to video copy detection that usually aims at detecting spatially and temporally transformed copies [7] and near-duplicates [9], video sequence identification rather aims at reliable and fast identification of identical or slightly different repetitions of video sequences of same temporal resolution. Döhring et al. [34] have recently presented a video identification approach based on gradient histograms that can quickly identify repetitions in TV broadcasts. However, their method is designed to find identical recurrences of content only (e.g. commercials) and does not work for slightly different content repetitions, e.g. such as music clips broadcasted by different TV channels with different content overlays (see Section 1).

3 Motion Signature

The video sequence identification approach presented in this paper is based on motion histograms. Our assumption is that motion in copies of original content is typically quite similar, even if some spatial transformations (e.g. resize, logo insertion, color change, etc.) have been applied to the copy.

We classify motion vectors of every frame into 12 different equidistant motion directions, each of 30 degrees. Therefore, we create one-frame-distance motion vectors predicted from frame $n-1$ to frame n (in presentation order). These one-frame-distance motion vectors of a particular frame are classified into a motion histogram with $K+1$ bins, modelling K equidistant motion direction intervals (bins $b \in \{1, \dots, K\}$) and a separate bin ($b=0$) for zero-length motion vectors,

as illustrated in Figure 3 for $K = 12$. We define the *direction* of a motion vector $\mu = (x, y) \neq (0, 0)$ with length $|\mu|$ as:

$$\omega(\mu) = \begin{cases} \arccos \frac{x}{|\mu|} & \text{if } y \geq 0 \\ 2\pi - \arccos \frac{x}{|\mu|} & \text{if } y < 0 \end{cases} \quad (1)$$

Note that $0 \leq \omega(\mu) < 2\pi$ and that $\omega(\mu) = 0$ corresponds to direction *right*. A motion vector μ is assigned to a bin of a $(K + 1)$ -bin motion histogram by the following equation:

$$b(\mu) = \begin{cases} 0 & \text{if } \mu = (0, 0) \\ 1 + (\lfloor \omega(\mu) \frac{K}{2\pi} + \frac{1}{2} \rfloor \bmod K) & \text{otherwise} \end{cases} \quad (2)$$

This formula has been chosen such that all motion directions differing from direction *right* ($\omega = 0$) by less than π/K (modulo 2π) are assigned to the same bin (bin 1). The same condition holds for the other main directions *left*, *up*, and *down*, too, if and only if K is a multiple of 4. Some more detailed information about our motion classification approach can be found in [8].

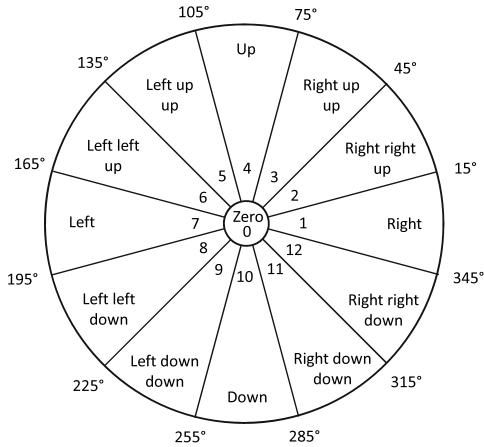


Fig. 3. Motion histogram representing the relative number of pixels moving to specific direction (or not moving at all (bin 0))

The motion vectors from frame $n - 1$ to frame n are estimated for small blocks of pixels (e.g. 4x4 pixels) using a Hexagon search [2][11] that favours short motion vectors and that is based on SAD measure. This motion estimation algorithm is typically also used by state-of-the-art video encoders, such as H.264/AVC [10]. Therefore, our approach is suited to directly use motion vectors contained in H.264/AVC compressed video streams. However, our evaluation videos recorded from DVB still use MPEG-2 as a video codec, as most of the DVB content in Europe, which performs motion estimation for 16x16 pixel blocks only. Through

experimental tests we found out that our video sequence identification approach, does not work very well with motion vectors estimated for such large blocks only. The motion histograms become much more distinctive when motion vectors for sub-partitions of macroblocks are available (e.g. 16x16, 16x8, ..., 4x4). Therefore, for the evaluations presented in this paper, we used the motion estimation algorithm of the open source H.264/AVC encoder x264 [1]. Note that this additional motion estimation step is not required if motion vectors are already available for small pixel areas (as e.g. for DVB content being already in H.264/AVC format).

4 Segment Matching Algorithm

Our video copy detection approach consists of two steps. In the first step it tries to find all matches of the signature $H(Q)$, of a given query Q (of length $|Q|$ frames), in the signature $H(V)$ of video stream V (of length $|V|$ frames). This is achieved by computing the Manhattan distance d_s between the query signature and the signature of the video stream at any possible starting frame s (see Equation B). The distance is computed over a sliding window W of a specific length that is typically a few seconds (in our evaluation we used 20 seconds). Any sequence S of length $|S| \geq 0.9 \times |Q|$ and a distance below a predefined threshold within the sliding window, i.e. $d_s \leq T$, is detected as a *match*.

Using a sliding window rather than a per-frame comparison makes our algorithm robust against short peaks of distance. On the other hand, it reduces the accuracy for detecting the correct start/stop position of a matching segment.

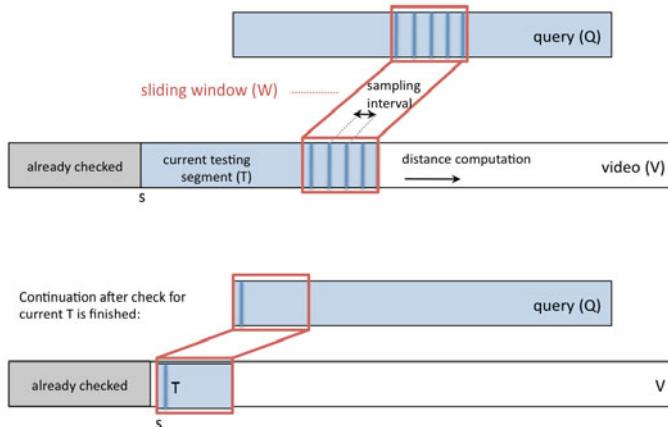
In order to speed up the search process we compute the distance for a specific sampling interval only (in our evaluation we used an interval of 25 frames). Figure 4 depicts the details of the entire search process. As shown in the figure, in either way - if the current testing segment matches or not - after the matching for the current T has finished, the search process continues with the next frame after s . This will typically produce several temporally overlapping hits for a single copy but enables to find the best matching segment (i.e. the segment with the lowest distance d_s). These "duplicate hits" are removed by the second step of our algorithm, which selects the hit with lowest distance only.

$$d_s = \sum_{i=|Q|-|W|}^{|Q|} \sum_{j=0}^{13} |H(Q)_{i,j} - H(V)_{s+i,j}| \quad (3)$$

$$\forall s = 0 \dots |V| - |Q|$$

5 Evaluation

We have recorded 48 hours of video from two music TV channels broadcasted via DVB-S in Europe (24 hours each) and performed the following video identification tasks:

**Fig. 4.** Copy detection process

- 1. Intra-Stream:** Given a query signature of a specific music clip from channel-1, find all other recurrences in channel-1.
- 2. Resized Intra-Stream:** Given a query signature of a specific music clip from channel-1, find all other recurrences in a copy of channel-1 having lower spatial resolution.
- 3. Inter-Stream:** Given a query signature of a specific music clip from channel-2, find all recurrences in channel-1 (whereas both recordings use the same spatial resolution).
- 4. Resized Inter-Stream:** Given a query signature of a specific music clip from channel-2, find all recurrences in a copy of channel-1 having a higher spatial resolution.

The recordings of both channels have been annotated manually in order to define ground truth. For all identification tasks we evaluated recall, precision, start/stop position accuracy, and run-time performance of our algorithm (without time required to estimate motion, see last paragraph of Section 3). We have evaluated our video sequence identification approach with several different values for threshold T within a meaningful range.

The 24 hours recording of channel-1 ("VIVA Germany") consists of music clips (55%), moderations/shows (3.47%), soaps (20.53%), and commercials (21%). In total the recording consists of 215 music clips, including repetitions. 165 music clips out of these 215 music clips are repetitions that consist of 54 unique music clips. Thus, in total there are 104 unique music clips in the 24 hours recording. Maximum repetition time is 7, average repetition time is 3.03, and minimum repetition time is 1. The maximum music clip length is 514 seconds, average length is 221 seconds, and minimum length is 42 seconds. The 24 hours recording of channel-2 ("MTV Germany") contains 21 unique music clips that appear 79 times in the recording of channel-1. For these inter-channel repetitions the maximum repetition time of a clip is 7, average repetition time is 3.76 and minimum repetition time is 2. All evaluations have been performed on a single

core of a quad-core CPU running at 2.68 GHz (Intel Core2Quad Q8300) with 4 GB RAM and Windows 7 64-bit.

5.1 Intra-Stream

In the *intra-stream* evaluation task we extracted the query signatures from the full-size video of channel-1 (768x576 pixel) and used them to find similar recurrences in the same video. For every group of recurrences (54 groups in total) we used that instance of a music clip as a query, which contained least transformations/overlays (i.e. the "clearest ones"). The results are shown in Table 1 and Figures 5 through 8.

Table 1. Results for *intra-stream* evaluation (T=threshold, R=recall, P=precision, Δ Start/Stop=secs missed at start/stop, RT=run-time in secs)

T	R	P	Δ Start	Δ Stop	RT
0.4	0.56	1.00	8.46	4.43	31.33
0.5	0.88	0.98	10.10	8.06	32.45
0.6	0.93	0.90	12.20	8.31	34.70
0.7	0.96	0.77	14.12	8.60	47.29
0.8	0.99	0.25	17.65	12.16	565.00

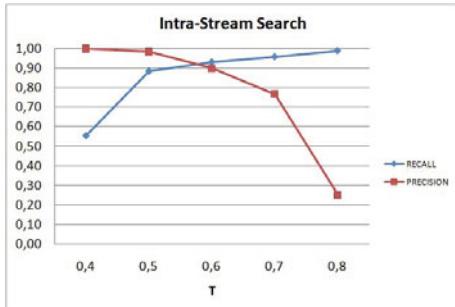


Fig. 5. Recall/Precision for *intra-stream* evaluation

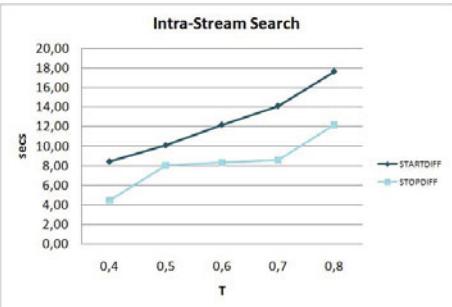


Fig. 6. Start/Stop accuracy for *intra-stream* evaluation

The recall/precision graph clearly shows that the optimal threshold setting for T is around 0.6. If the threshold is below 0.5 only a few (typically very similar or equal) repetitions are detected, which results in a low recall. Vice versa, if the threshold is above 0.7, recall increases but also precision strongly decreases because too much frames match, which yields in many false positives. As shown in Figure 6, the higher the threshold T the higher is also the average inaccuracy for detecting the start and stop positions. With $T = 0.6$ we have an average start inaccuracy of 12.2 secs, which is 5.52 % of average music clip length. However, please note that average start/stop accuracy can never be zero for our evaluation data, since repetitions of the same music clip are often of different length.

5.2 Resized Intra-Stream

In the *resized intra-stream* evaluation we extracted the query signatures from the full-size video of channel-1 (768x576 pixel) and used them to find similar recurrences in a quarter-size copy (384x288 pixel) of the same video. The results are shown in Table 2 and Figures 7 through 8.

Table 2. Results for *resized intra-stream* evaluation (T =threshold, R =recall, P =precision, Δ Start/Stop=secs missed at start/stop, RT=run-time in secs)

T	R	P	Δ Start	Δ Stop	RT
0.4	0.18	0.71	26.55	14.39	33.90
0.5	0.48	0.69	34.82	19.65	33.82
0.6	0.70	0.65	28.28	14.34	36.03
0.7	0.87	0.61	24.67	14.24	44.93
0.8	0.90	0.39	27.26	16.65	239.78

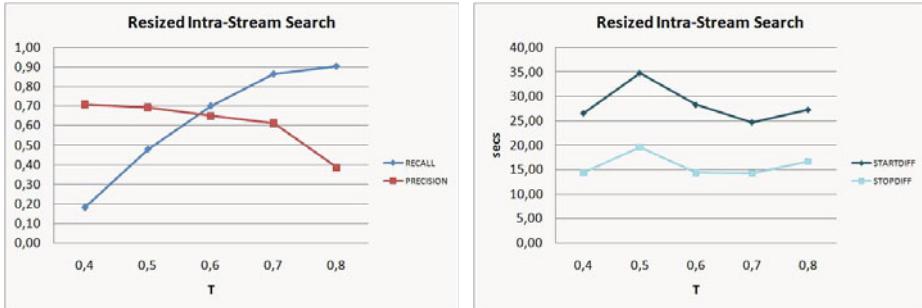


Fig. 7. Recall/Precision for *resized intra-stream* evaluation **Fig. 8.** Start/Stop accuracy for *resized intra-stream* evaluation

These results show that our video sequence identification approach basically works quite well even for videos with a high spatial difference (the testing video was only $\frac{1}{4}$ size of the query video): with $T = 0.7$ we can still find almost 90 percent of recurrences, although with a moderate number of false positives. However, the recall and precision values are significantly lower than for the first evaluation, where we had test and query video of same spatial resolution.

In Figure 8 we can see that the average start/stop accuracy is highly dependent on the recall. The reason for the peak from $T = 0.4$ to $T = 0.5$ is the very low recall at $T = 0.4$ ($R=0.18$), which means that the computation of start/stop accuracy is based on a few obviously accurate matches only.

5.3 Inter-Stream

In the *inter-stream* evaluation task we evaluated how good our approach performs at detecting repetitions across different TV broadcasting channels. More

precisely, we extracted the query signatures from the full-size video of channel-2 (768x576 pixel) and used them to find similar recurrences in channel-1 of same spatial resolution. For each of the 21 groups of different music clips in channel-2 that also appear in channel-1, we used a random instance as a query and evaluated the average performance. The results are shown in Table 3 and Figures 9 through 10. As these results show, in this task we achieve quite high values for recall and precision: with $T = 0.6$ both are near to 1.0. However, the reason why this test case performed better than the first one (intra) is simply that channel-2 does not use such large areas for overlays than channel-1 does. As a consequence we can say that the less distortions are contained in the query video sequence, the better our video sequence identification approach works.

Table 3. Results for *inter-stream* evaluation (T =threshold, R =recall, P =precision, Δ Start/Stop=secs missed at start/stop, RT=run-time in secs)

T	R	P	Δ Start	Δ Stop	RT
0.4	0.31	1	25.54	7.50	29.34
0.5	0.76	1	23.44	9.42	30.92
0.6	0.99	0.975	22.32	9.95	34.92
0.7		1	0.84	22.02	9.72
0.8		1	0.33	22.02	9.72
					81.20

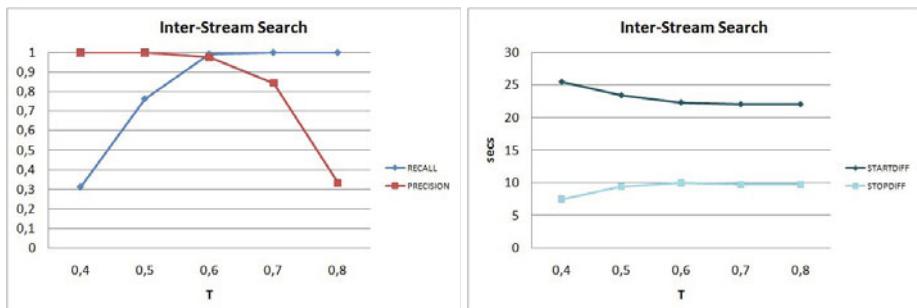


Fig. 9. Recall/Precision for *inter-stream* evaluation **Fig. 10.** Start/Stop accuracy for *inter-stream* evaluation

5.4 Resized Inter-Stream

For this evaluation we used query signatures extracted from motion histograms of a quarter-size video copy (384x288 pixel) of channel-2 to find similar recurrences in the full-size video (768x576 pixel) of channel-1. This test case used the same queries as the previous test case, however, the signatures were different. The results are shown in Table 4 and Figures 11 through 12 (results for threshold setting $T = 0.4$ have been omitted since not a single match was found).

In comparison to the resized intra-stream evaluation the evaluation results of this test case are much better: with $T = 0.6$ we achieve recall and precision above 90% although our query video was only of $\frac{1}{4}$ spatial size. This proves our observation in the previous test case that the less transformed the content from which the query signature is extracted, the better the detection performance.

Table 4. Results for *resized inter-stream* evaluation (T =threshold, R =recall, P =precision, Δ Start/Stop=secs missed at start/stop, RT=run-time in secs)

T	R	P	Δ Start	Δ Stop	RT
0.5	0.40	1.00	24.84	8.98	41.24
0.6	0.95	0.98	25.97	9.95	41.18
0.7	0.98	0.87	25.22	9.68	51.85
0.8	0.99	0.41	24.20	9.84	85.78

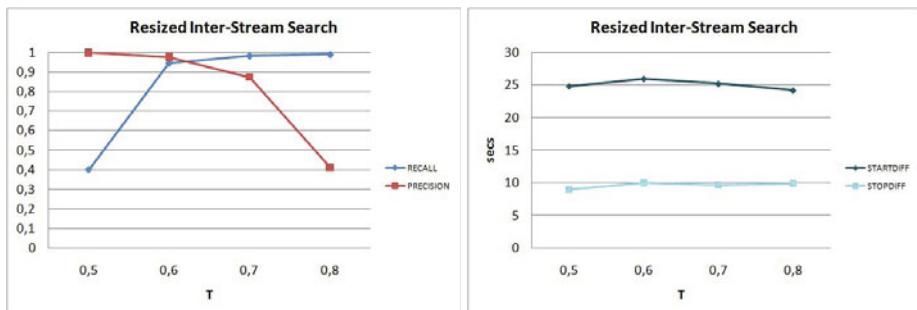


Fig. 11. Recall/Precision for *resized inter-stream* evaluation **Fig. 12.** Start/Stop accuracy for *resized inter-stream* evaluation

6 Conclusion

We have presented a video sequence identification approach that can reliably detect recurring video segments in TV broadcasts. Our method uses motion vector classification and is based on the idea that identical or highly similar content has highly similar motion flows. The presented approach works for both repetitions of same content and repetitions of slightly different content. In our evaluations with 48 hours of video we have shown that our method works well and achieves high recall values together with a very low number of false positives. The presented method has low run-time requirements and works in a fraction of real-time, which makes it suitable for on-the-fly monitoring of TV broadcasts in order to reliably detect content repetitions.

References

1. Aimar, L., Merritt, L., Petit, E., Chen M., Clay, J., Rullgard, M., Czyz, R., Heine, C., Izvorski, A., Wright, A.: x264 - a free H264/AVC encoder (2010),
<http://www.videolan.org/developers/x264.html> (last accessed on: 13/07/10)
2. Chen, Z.: Efficient block matching algorithm for motion estimation. International Journal of Signal Processing 5(2), 133–137 (2009)
3. Döhring, I., Lienhart, R.: Mining tv broadcasts for recurring video sequences. In: Proceeding of the ACM International Conference on Image and Video Retrieval, CIVR 2009, pp. 1–8. ACM, New York (2009)
4. Ohring, I.D., Lienhart, R.: Mining TV Broadcasts 24/7 for Recurring Video Sequences. In: Video Search and Mining, pp. 327–356 (2010)
5. Hampapur, A., Hyun, K., Bolle, R.: Comparison of sequence matching techniques for video copy detection. In: Conference on Storage and Retrieval for Media Databases, pp. 194–201. Citeseer (2002)
6. Law-To, J., Chen, L., Joly, A., Laptev, I., Buisson, O., Gouet-Brunet, V., Boujemaa, N., Stentiford, F.: Video copy detection: a comparative study. In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval, p. 378. ACM, New York (2007)
7. Liu, Z., Liu, T., Gibbon, D., Shahraray, B.: Effective and scalable video copy detection. In: Proceedings of the International Conference on Multimedia Information Retrieval, pp. 119–128. ACM, New York (2010)
8. Schoeffmann, K., Lux, M., Taschwer, M., Boeszoermenyi, L.: Visualization of video motion in context of video browsing. In: Proceedings of the IEEE International Conference on Multimedia and Expo, New York, USA, IEEE, Los Alamitos (July 2009)
9. Tan, H., Ngo, C., Hong, R., Chua, T.: Scalable detection of partial near-duplicate videos by visual-temporal consistency. In: Proceedings of the Seventeen ACM International Conference on Multimedia, pp. 145–154. ACM, New York (2009)
10. Wiegand, T., Sullivan, G.J., Bjontegaard, G., Luthra, A.: Overview of the H.264/AVC Video Coding Standard. IEEE Transactions on Circuits and Systems for Video Technology (July 2003)
11. Zhu, C., Lin, X., Chau, L.P.: Hexagon-based search pattern for fast block motion estimation. IEEE Transactions on Circuits and Systems for Video Technology 12(5), 349 (2002)

Content-Based Multimedia Retrieval in the Presence of Unknown User Preferences

Christian Beecks¹, Ira Assent², and Thomas Seidl¹

¹ Data Management and Data Exploration Group

RWTH Aachen University, Germany

{beecks, seidl}@cs.rwth-aachen.de

² Department of Computer Science

Aarhus University, Denmark

ira@cs.au.dk

Abstract. Content-based multimedia retrieval requires an appropriate similarity model which reflects user preferences. When these preferences are unknown or when the structure of the data collection is unclear, retrieving the most preferable objects the user has in mind is challenging, as the notion of similarity varies from data to data, from task to task, and ultimately from user to user. Based on a specific query object and unknown user preferences, retrieving the most similar objects according to some default similarity model does not necessarily include the most preferable ones. In this work, we address the problem of content-based multimedia retrieval in the presence of unknown user preferences. Our idea consists in performing content-based retrieval by considering all possibilities in a family of similarity models simultaneously. To this end, we propose a novel content-based retrieval approach which aims at retrieving all potentially preferable data objects with respect to any preference setting in order to meet individual user requirements as much as possible. We demonstrate that our approach improves the retrieval performance regarding unknown user preferences by more than 57% compared to the conventional retrieval approach.

Keywords: content-based retrieval, user preferences, multimedia databases.

1 Introduction

Content-based multimedia retrieval [4, 8, 14, 18, 19] supports users in accessing and searching voluminous multimedia collections comprising image, audio, video, or other non-text data. Providing a suitable similarity model which captures the user's notion of similarity, content-based multimedia retrieval systems are able to return the most similar objects with respect to a specified query object effectively and efficiently.

For this purpose, the similarity model captures the multimedia objects' inherent features describing the object contents as well as a similarity measure defining the similarity between two feature representations. As the notion of



Fig. 1. Example of different user preferences causing different retrieval results

similarity highly depends on the context and thus varies from data to data, from task to task, and ultimately from user to user, the similarity model's properties are frequently expressed by means of so-called *user preferences* which reflect the user's notion of similarity in the current context. In this way, including user preferences by modifying the multimedia retrieval system's similarity measure influences the content-based retrieval process and its results. Figure II depicts an example where different images are returned to a user specifying different user preferences. Both rows depict the content-based retrieval results of a similarity query with the query object depicted on the left-hand side. While the resulting images shown in the top row share the same color and shape, the ones shown in the bottom row basically agree on a similar shape.

This example indicates that different user preferences adapting the similarity measure and/or the feature representation will lead to different retrieval results and that the quality of the results also depends on the user preferences. If the similarity model, adapted to the user preferences, reflects the user's notion of similarity, the most similar multimedia objects returned to the user are also the most preferable ones.

Let us now consider the case of unspecified or even unknown user preferences. In this case, retrieving the most preferable objects the user has in mind is challenging. Reconsider the example above, if a user is looking for flowers of similar color, then results of similar shape are not relevant.

In practice, content-based retrieval typically assumes some default similarity model. Nevertheless, by neglecting user preferences, the probability that the retrieved results correspond to the user's notion of similarity and contain preferable data objects is low. The most similar multimedia objects provided by the content-based retrieval system when relying on a default similarity model are not necessarily the most preferable ones the user has in mind.

In this work, we address the problem of content-based multimedia retrieval supporting incomplete, unspecified, or even unknown user preferences. Our idea consists in performing content-based retrieval by considering all possible user preferences within a family of similarity models simultaneously in order to gather all potentially preferable data objects. To this end, we propose a novel content-based retrieval approach which aims at retrieving all potentially preferable data objects with respect to any preference setting in order to meet individual user

requirements as much as possible. We claim that only by providing all the best options users are able to make truly informed choices.

Our contributions include:

- A novel content-based multimedia retrieval model supporting unknown user preferences within any kind of similarity model.
- An example for the particular case of weighted Euclidean similarity model showing how to retrieve all preferable data objects efficiently.
- An evaluation on benchmark image data showing that our approach maximizes the number of potentially preferable data objects.

This paper is structured as follows: in Section 2, we review related work on content-based multimedia retrieval. Section 3 introduces our general retrieval approach, discusses its main characteristics, and investigates an example for the particular case of the weighted Euclidean similarity model. Section 4 evaluates the retrieval performance of our approach regarding unknown user preferences. Finally, we conclude this work in Section 5 and highlight future work.

2 Related Work

In multimedia retrieval, the question on how to adequately represent user preferences has long since been an active area of research. Much research has focused on defining and evaluating suitable similarity models [16,10]. It is widely accepted that similarity is subjective and context-dependent. As a consequence, different approaches for fine-tuning similarity or providing an overview of the data have been studied [9,11,22]. Existing approaches, however, assume that the context and user preferences are either known in advance or are learned in an iterative process, as e.g. in relevance feedback [7,23]. This is problematic when the user is not aware of the content of the multimedia collection and hence cannot understand the effect of different similarity settings given by his or her user preferences.

While content-based multimedia retrieval [4,8,14,18,19] is well examined, finding all the best options has been studied as the *maximum vector problem*, the *floating currency problem*, or *Pareto optimality* in computational geometry or economics [12,13,16]. In the database field, the *skyline* operator with different models and algorithms has been researched recently [3]. The skyline is defined as the set of objects which are not *dominated* by other objects, i.e. which are the best options with respect to some preference. Thus, the set contains the optimum with respect to arbitrary monotonous scoring functions for all attributes. Papadias et al. [15] introduced the dynamic skyline, i.e. a skyline not in absolute terms (origin of the feature space), but with respect to a query object. This is further studied in [17,20].

In this work, we claim that content-based multimedia retrieval in the presence of unknown user preferences is akin to the problem of finding all the best options which are given by retrieving multimedia objects with respect to all possible preference settings. In this case, different user preferences correspond to different

optima, i.e. relevant objects, and the ideal retrieval should gather all potentially preferable data objects at once. We investigate this issue in the following section where we present our novel retrieval approach.

3 A Novel Retrieval Approach

In this section, we introduce our novel content-based retrieval approach. For this purpose, we first formalize the conventional content-based retrieval process and then introduce our *unknown preference retrieval model* which aims at retrieving all potentially preferable data objects with respect to any preference setting. Finally, we explain how to efficiently process queries for the case of the weighted Euclidean distance.

Content-based multimedia retrieval is performed by specifying a *similarity model* comprising the feature representation, similarity measure, and user preferences. In the retrieval process, data objects which exhibit the highest similarity values with respect to a given query object are returned to the user. While the feature representation, such as feature vectors or feature signatures [2], digitizes and compactly stores the data objects' inherent properties [5], the similarity measure, such as those evaluated in [110], computes similarity values among the feature representations of the query object and each database object. User preferences are frequently incorporated by weighting the similarity measure in order to match the user's notion of similarity. We define the general similarity model involving the aforementioned aspects below.

Definition 1. *Similarity model*

Given a universe \mathcal{U} of multimedia objects, a feature representation $f : \mathcal{U} \rightarrow \mathcal{FS}$, and a similarity measure $sim_\phi : \mathcal{FS} \times \mathcal{FS} \rightarrow \mathcal{R}^{\geq 0}$ with user preferences ϕ , the similarity model $S : \mathcal{U} \times \mathcal{U} \rightarrow \mathcal{R}^{\geq 0}$ is defined as:

$$S(q, p) \mapsto sim_\phi(f(q), f(p)).$$

According to Definition 1, the similarity model S measures the similarity between two multimedia objects q and p by evaluating the corresponding similarity measure sim_ϕ on the multimedia objects' feature representations $f(q)$ and $f(p)$. Without loss of generality, we thus assume that the user preferences ϕ only adapt the parameters of the similarity measure sim_ϕ .

Based on the definition of the similarity model S , we define the result of a content-based retrieval process as a *ranking* of the multimedia collection according to the query and the specified similarity model as in the following definition.

Definition 2. *Ranking*

Given a collection $o_1, \dots, o_n \in \mathcal{M} \subseteq \mathcal{U}$ of multimedia objects, a similarity model S , and a query object $q \in \mathcal{U}$, the resulting ranking $R(q, \mathcal{M})$ of the retrieval process is defined as:

$$R(q, \mathcal{M}) := (o_{\pi(1)}, \dots, o_{\pi(n)}),$$

where $o_{\pi(i)}$ denotes the multimedia object at position i , and it holds that $S(q, o_{\pi(i)}) \geq S(q, o_{\pi(j)})$ for all data objects $o_{\pi(i)}$ and $o_{\pi(j)}$ with $i < j$.

The ranking $R(q, \mathcal{M})$ of the retrieval process sorts the data objects of the multimedia collection \mathcal{M} according to their similarities to the query object q . The most similar objects are at the head of the ranking and, thus, they are the most preferable ones given a specific similarity model. The definition of the *most preferable objects model* is given below.

Definition 3. Most preferable objects model (MPO)

Given a collection $o_1, \dots, o_n \in \mathcal{M} \subseteq \mathcal{U}$ of multimedia objects, a similarity model S , and a query object $q \in \mathcal{U}$, the set of most preferable objects $R_k^*(q, \mathcal{M})$ of minimum size k is recursively defined as:

$$R_0^*(q, \mathcal{M}) := \emptyset,$$

$$R_k^*(q, \mathcal{M}) := \{o \in \mathcal{M} \mid \forall o' \in \mathcal{M} - R_{k-1}^*(q) : S(q, o') \leq S(q, o)\}.$$

According to Definition 3, the set $R_k^*(q, \mathcal{M})$ contains k most preferable data objects, i.e. the k nearest neighbors, which match the user preferences as much as possible.¹ It depends on the specified similarity model and the query object. The properties of the similarity model are adapted to individual user preferences. For this reason, the most similar multimedia objects are the most preferable ones, because it is assumed that the users' similarity notions are completely reflected by their preferences.

However, if the user preferences are unspecified or even unknown, the set of most preferable data objects is not defined. Simply assuming some default similarity model may result in irrelevant results although the returned multimedia objects agree on some kind of default similarity with the query object. We state this issue of *unknown user preferences* as below.

Problem 1. Unknown user preferences

Given a collection $\mathcal{M} \subseteq \mathcal{U}$ of multimedia objects, a similarity model S with unknown user preferences ϕ , and a query object $q \in \mathcal{U}$, it is unclear which objects $o \in \mathcal{M}$ are the most preferable ones and thus have to be returned to the user.

As a solution we propose retrieving all potentially preferable data objects contained in the multimedia collection by taking into account all possible user preferences. This ensures that the retrieval system covers all user preferences and returns a diverse set of data objects which contains the most preferable object the user has in mind. To this end, we define the *unknown preference retrieval model* in the following definition.

Definition 4. Unknown preference retrieval model (UPR)

Given a collection $\mathcal{M} \subseteq \mathcal{U}$ of multimedia objects, a similarity model S with unknown user preferences ϕ , and a query object $q \in \mathcal{U}$, the resulting set $R_U(q, \mathcal{M})$ of the retrieval process is defined as:

¹ The cardinality of the set $R_k^*(q, \mathcal{M})$ is exactly k if we assume no ambiguity of similarity values among query and data objects, i.e. $\forall o, o' \in \mathcal{M} : S(q, o) \neq S(q, o')$, otherwise $R_k^*(q, \mathcal{M})$ can contain more than k objects.

$$R_U(q, \mathcal{M}) := \{o \in \mathcal{M} \mid o \text{ is the most preferable object w.r.t. some user preference } \phi\}.$$

According to Definition 4, the result set $R_U(q, \mathcal{M})$ of the unknown preference retrieval model contains all data objects for which there exist some user preferences making this object the most preferable one. In other words, this set is complete with respect to any user preference setting. Any data object not contained in $R_U(q, \mathcal{M})$ is not the most preferable object for any user preference. We state the completeness of $R_U(q, \mathcal{M})$ in the following lemma.

Lemma 1. *Completeness w.r.t. any user preferences*

Given a collection $\mathcal{M} \subseteq \mathcal{U}$ of multimedia objects, a similarity model S with unknown user preferences ϕ , and a query object $q \in \mathcal{U}$, the following holds:

$$o \notin R_U(q, \mathcal{M}) \Leftrightarrow \forall \phi : o \notin R_1^*(q, \mathcal{M}).$$

Proof. According to Definition 4, each data object $o \in R_U(q, \mathcal{M})$ is the most preferable object for some user preference ϕ , i.e. $o \in R_U(q, \mathcal{M}) \Leftrightarrow \exists \phi : o \in R_1^*(q, \mathcal{M})$. Consequently, by inverting this formula, we obtain the statement.

So far, we have theoretically introduced and formalized our novel content-based retrieval approach which aims at retrieving all potentially preferable objects in the presence of unknown user preferences. In the remainder of this section, we will elucidate our approach based on the weighted Euclidean similarity model and describe how to implement our approach efficiently by making use of dynamic skyline algorithms. For this purpose, we suppose the feature representation $f(o) = \mathbf{o} \in \mathcal{FS} \subseteq \mathcal{R}^n$ to be of the form of an n -dimensional vector, for instance a color histogram, and the similarity measure $sim_\phi(f(p), f(q)) = \sqrt{\sum_{i=1}^n \phi_i \cdot (p_i - q_i)^2}$ to be the weighted Euclidean distance with weights $\phi_i \in \mathcal{R}^+$ reflecting user preferences.

Given this similarity model, we illustrate the principle of the *most preferable objects model* $R_k^*(q, \mathcal{M})$ compared with that of the *unknown preference retrieval model* $R_U(q, \mathcal{M})$ in Figure 2 where we depict the retrieved data objects as blue points. As one can see in this example, the conventional retrieval approach yields the sets of preferable objects $R_1^*(q, \mathcal{M}) \subseteq R_2^*(q, \mathcal{M}) \subseteq R_3^*(q, \mathcal{M}) = \{o_1, o_2, o_3\}$ covering only a small region within the multimedia collection. Therefore these sets also cover only a small range of possible user preferences. Increasing the number $k = 1 \dots 3$ will not necessarily cover a broader range of user preferences. In contrast, this is achieved by making use of the unknown preference retrieval model which returns the set $R_U(q, \mathcal{M}) = \{o_1, o_4, o_5\}$ comprising all preferable data objects with respect to the given query q . In this way, our approach considers the results as a set of data objects covering all possible user preferences rather than a sorted ranking corresponding to a restricted amount of user preferences.

In order to compute and retrieve this set $R_U(q, \mathcal{M})$ in a feasible amount of time without evaluating all possible user preferences, we make use of the fact that this problem computationally corresponds to a dynamic skyline problem [15],

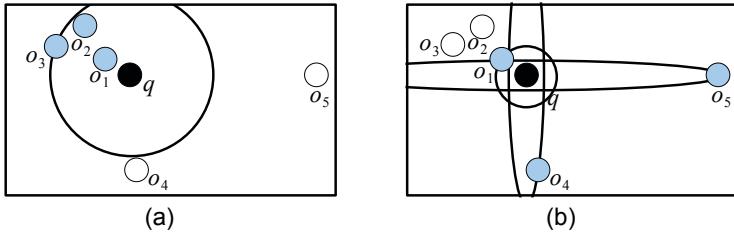


Fig. 2. Comparison of (a) the *most preferable objects model* $R_{k=3}^*(q, \mathcal{M}) = \{o_1, o_2, o_3\}$ and (b) the *unknown preference retrieval model* $R_U(q, \mathcal{M}) = \{o_1, o_4, o_5\}$

cf. Section 2. The result set contains all data objects that are most preferable with respect to some user preference and with respect to a query object, i.e. it corresponds to the dynamic skyline which originates at the query point given by the user. In the database community, several algorithms have been proposed: In a branch-and-bound approach [15], a heap is used to process data objects with respect to their distance from the dynamic query object in a best-first fashion. More recently, [17] propose a caching mechanism to use previous dynamic queries to prune future ones. As a result of dynamic skyline algorithms, the returned data objects, contained in the skyline, are optimal in terms of a monotonic scoring function optimizing a set of attributes which are user preferences ϕ in our particular case. Thus, we are able to retrieve the set $R_U(q, \mathcal{M})$ efficiently by making use of existing skyline algorithms.

In the next section, we present experimental results.

4 Experimental Evaluation

In this section, we compare the retrieval performance for unknown user preferences of our approach, *unknown preference retrieval model* (UPR), with that of the conventional retrieval approach, *most preferable objects model* (MPO), based on the *Corel Wang* database [21] comprising 1,000 images from 10 different themes. For this purpose, we extracted 30-dimensional feature histograms including position, color, and texture information [5] and used weighted Euclidean distance function with weights $\phi_i \in \mathcal{R}^+$ for $1 \leq i \leq 30$ reflecting user preferences, as described in the previous section. We used objects from each theme as queries and randomly determined a subset of unknown weights ϕ_i of the corresponding size. By repeating this query process 50 times for each data object, we averaged the number of different most preferable objects observed in the themes.

In Figure 3 we depict the average number of different most preferable objects w.r.t. some user preferences ϕ per theme of the *Corel Wang* database by varying the number of unknown weights ϕ_i between 10 and 30. The depicted values indicate the minimum number of different user preferences hidden in the data, as each most preferable object covers at least one user preference. However, it is

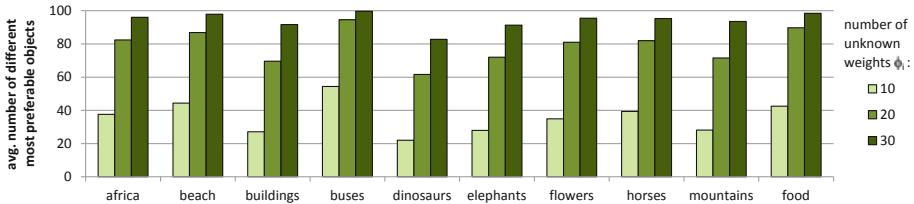


Fig. 3. Average number of different most preferable objects w.r.t. some user preferences ϕ by varying the number of unknown weights ϕ_i between 10 and 30

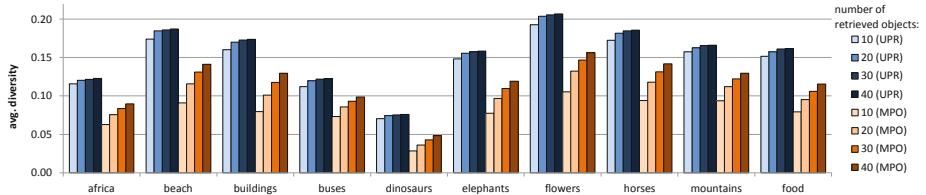


Fig. 4. Average *diversity* of the retrieval results by varying the number of retrieved data objects between 10 and 40

more likely that there exists some most preferable objects which cover an infinite number of user preferences. As can be seen in the figure, the number of different most preferable objects varies from theme to theme and additionally depends on the number of unknown weights ϕ_i . While the number of most preferable objects reaches its minimum value of 22 in the *dinosaurs* theme, it reaches the maximum value of 99 in the *buses* theme. On average, we observed 35, 79, and 94 different most preferable objects by setting the number of unknown weights to 10, 20, and 30, respectively. This indicates the challenge of maximizing the number of potentially preferable objects returned to the users in the presence of unknown user preferences, as this number is typically significantly higher than the number of retrieved data objects.

In order to evaluate the retrieval performance of our and the conventional retrieval approach, we measured the average *diversity* of the retrieval results of both approaches. The *diversity* is defined as expected value of the distance distribution of the retrieval results $R'(q, \mathcal{M})$:

$$\text{diversity}(R'(q, \mathcal{M})) = \sum_{o, o' \in R'(q, \mathcal{M})} p(L_2(f(o), f(o'))) \cdot L_2(f(o), f(o')),$$

where we assume a uniform distance probability distribution. We argue that homogeneous retrieval results will become more frequent in a compact region in the feature space while heterogeneous ones will be more scattered in the feature space. Thus, a high expected value of the distance distribution reflects high diversity of the retrieval results and vice versa. The number of retrieved data objects varied between 10 and 40 and for the unknown preference retrieval approach (UPR) this number was obtained by performing a random sampling

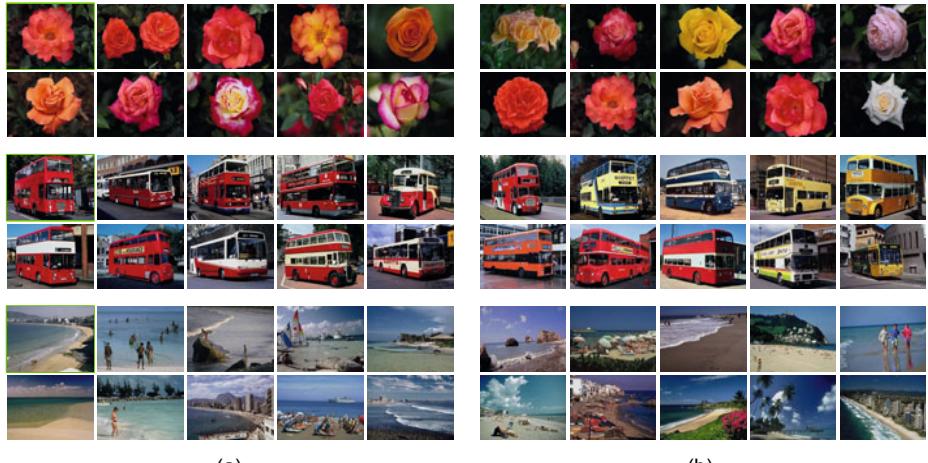


Fig. 5. Retrieval results of (a) the conventional retrieval approach (MPO) and (b) the unknown preference retrieval approach (UPR)

of elements from the set $R_U(q, \mathcal{M})$. These results are averaged over 1,000 randomized queries by varying the number of unknown weights ϕ_i between 10 and 30. In Figure 4 we visualize the results of our proposed approach (UPR) and the conventional retrieval approach (MPO) with bluish and orangish color, respectively. As can be seen in the figure, the average diversity of the conventional retrieval approach (MPO) lies between 0.029 and 0.156 depending on the size of the retrieval set $R_k^*(q, \mathcal{M})$. In contrast, the average diversity of the unknown preference retrieval approach (UPR) lies between 0.070 and 0.207, also depending on the size of the set $R_U(q, \mathcal{M})$. As a result, our approach increases the diversity and maximizes the number of most preferable objects returned to the user. On average, it improves the retrieval performance by more than 57%.

We complete the experimental evaluation by depicting some retrieval results in Figure 5 where we visualize the retrieved images of three different queries of (a) the conventional retrieval approach (MPO) and (b) the proposed unknown preference retrieval approach (UPR). The results of the conventional retrieval approach which are depicted on the left-hand side are very homogeneous, i.e. the images are visually similar to the query image (top left, green border). Consequently, the number of preferable objects gathered by the conventional retrieval approach is limited, as similar images frequently cover similar user preferences. The diversity and thus heterogeneity of the retrieved images is increased on the right-hand side. By using the proposed unknown preference retrieval approach the similarity of the images decreases, meaning higher diversity regarding the number of preferable objects returned to the user.

To sum up, we have shown that content-based retrieval in the presence of unknown user preferences is a difficult and challenging task. While conventional retrieval approaches do not include all potentially preferable multimedia objects in the results, our approach incorporates all possible user preferences and thus gathers all preferable multimedia objects at once.

5 Conclusions and Future Work

We presented a novel content-based retrieval approach supporting unknown user preferences. By retrieving all potentially preferable data objects with respect to any preference setting, our approach advances in meeting individual user requirements as much as possible and improves the retrieval performance in terms of diversity by more than 57%.

We believe that this approach opens up new directions for research, most specifically in devising measures for the diversity of retrieval results that reflect user perception. Also, for applications where the result size should be fixed to a small set, finding a diverse representative result set is an interesting question. In order to retrieve a representative set, we intend to investigate different techniques and heuristics which maximize the diversity according to possible user preferences. In this manner, a relevance feedback style mechanism could be employed to guide the users to their preferred objects. Finally, we intend to conduct user studies to evaluate how well user preferences are explored and met.

Acknowledgments. The projects underlying this report were funded by the German Federal Ministry of Economics and Technology under project funding reference number 01MQ09014 and by the Deutsche Forschungsgemeinschaft (DFG) within the Collaborative Research Center (SFB) 686. The responsibility for the content of this publication lies with the authors.

References

1. Beecks, C., Uysal, M.S., Seidl, T.: A comparative study of similarity measures for content-based multimedia retrieval. In: Proceedings of the IEEE International Conference on Multimedia & Expo., pp. 1552–1557 (2010)
2. Beecks, C., Uysal, M.S., Seidl, T.: Signature quadratic form distance. In: Proceedings of the ACM International Conference on Image and Video Retrieval, pp. 438–445 (2010)
3. Börzsönyi, S., Kossmann, D., Stocker, K.: The Skyline operator. In: Proceedings of the IEEE International Conference on Data Engineering, pp. 421–430 (2001)
4. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image Retrieval: Ideas, Influences, and Trends of the New Age. ACM Computing Surveys 40(2), 1–60 (2008)
5. Deselaers, T., Keysers, D., Ney, H.: Features for Image Retrieval: An Experimental Comparison. Information Retrieval 11(2), 77–107 (2008)
6. Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W.: Efficient and Effective Querying by Image Content. Journal of Intelligent Information Systems 3(3/4), 231–262 (1994)
7. Ferreira, C., Santos, J., da, S., Torres, R., Gonçalves, M., Rezende, R., Fan, W.: Relevance feedback based on genetic programming for image retrieval. Pattern Recognition Letters (in Press, 2010)
8. Geetha, P., Narayanan, V.: A Survey of Content- Based Video Retrieval. Journal of Computer Science 4(6), 474–486 (2008)
9. Heesch, D.: A survey of browsing models for content based image retrieval. Multimedia Tools and Applications 40(2), 261–284 (2008)

10. Hu, R., Rüger, S.M., Song, D., Liu, H., Huang, Z.: Dissimilarity measures for content-based image retrieval. In: Proceedings of the IEEE International Conference on Multimedia & Expo., pp. 1365–1368 (2008)
11. Ishikawa, Y., Subramanya, R., Faloutsos, C.: Mindreader: Querying databases through multiple examples. In: Proceedings of the International Conference on Very Large Data Bases, pp. 218–227 (1998)
12. Katzner, D.W.: An introduction to the economic theory of market behavior: microeconomics from a Walrasian perspective. Edward Elgar Publishing (2008)
13. Kung, H.T., Luccio, F., Preparata, F.P.: On finding the maxima of a set of vectors. *Journal of the ACM* 22(4), 469–476 (1975)
14. Lew, M.S., Sebe, N., Djerafa, C., Jain, R.: Content-Based Multimedia Information Retrieval: State of the Art and Challenges. *ACM Transactions on Multimedia Computing, Communications and Applications* 2(1), 1–19 (2006)
15. Papadias, D., Tao, Y., Fu, G., Seeger, B.: An optimal and progressive algorithm for skyline queries. In: Proceedings of the ACM International Conference on Management of Data, pp. 467–478 (2003)
16. Preparata, F.P., Shamos, M.I.: Computational Geometry: An Introduction. Springer, Heidelberg (1985)
17. Sacharidis, D., Bouros, P., Sellis, T.: Caching dynamic skyline queries. In: Ludäscher, B., Mamoulis, N. (eds.) SSDBM 2008. LNCS, vol. 5069, pp. 455–472. Springer, Heidelberg (2008)
18. Sebe, N., Lew, M.S., Zhou, X., Huang, T.S., Bakker, E.M.: The state of the art in image and video retrieval. In: Proceedings of the 2nd International Conference on Image and Video Retrieval, pp. 7–12 (2003)
19. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12), 1349–1380 (2000)
20. Tan, K.L., Eng, P.K., Ooi, B.C.: Efficient progressive skyline computation. In: Proceedings of the International Conference on Very Large Data Bases, pp. 301–310 (2001)
21. Wang, J., Li, J., Wiederhold, G.: SIMPLICITY: semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(9), 947–963 (2001)
22. Wichterich, M., Beecks, C., Sundermeyer, M., Seidl, T.: Exploring multimedia databases via optimization-based relevance feedback and the earth mover’s distance. In: Proceedings of the ACM Conference on Information and Knowledge Management, pp. 1621–1624 (2009)
23. Zhou, X., Huang, T.: Relevance feedback in image retrieval: A comprehensive review. *Multimedia systems* 8(6), 536–544 (2003)

People Localization in a Camera Network Combining Background Subtraction and Scene-Aware Human Detection

Tung-Ying Lee¹, Tsung-Yu Lin¹, Szu-Hao Huang¹,
Shang-Hong Lai¹, and Shang-Chih Hung²

¹ Department of Computer Science, National Tsing Hua University,
HsinChu 300, Taiwan

² Identification and Security Technology Center,
Industrial Technology Research Institute, Taiwan

Abstract. In a network of cameras, people localization is an important issue. Traditional methods utilize camera calibration and combine results of background subtraction in different views to locate people in the three dimensional space. Previous methods usually solve the localization problem iteratively based on background subtraction results, and high-level image information is neglected. In order to fully exploit the image information, we suggest incorporating human detection into multi-camera video surveillance. We develop a novel method combining human detection and background subtraction for multi-camera human localization by using convex optimization. This convex optimization problem is independent of the image size. In fact, the problem size only depends on the number of interested locations in ground plane. Experimental results show this combination performs better than background subtraction-based methods and demonstrate the advantage of combining these two types of complementary information.

Keywords: Probabilistic occupancy map, video surveillance, human localization, multi-camera surveillance.

1 Introduction

To locate people in a network of cameras is an important issue in video surveillance environments or sport grounds. Locations of people can be utilized to track humans [1], to analyze sport players, to recognize them [2], and to perform behavior analysis [3].

Traditional human detectors are designed for detecting pedestrians from a single view. Recently, several features have been used in human detection. Local edge orientation histograms (EOH) [4] and region covariance features [5] are used. Especially, Histograms of Oriented Gradient (HOG) [6] are very popular and proved to be quite successful for application to human detection. The conventional Haar-like features are also used in human detectors [7]. The Adaboost framework can be used for combining these different features [5, 7]. However, partial occlusion is a difficult problem

that severely degrades the performance of human detectors. In spite that some researchers exploit information on manifold to handle partial occlusion [8], the single view approach is still not suitable for surveillance environment with multiple cameras. In a network of cameras, information gathered from multiple well-set cameras will compensate the partial occlusion in single view. Hence, background subtraction-based human detection is reasonable in a network of cameras.

Background subtraction-based approach can produce an occupancy map. In most situations, a ground plane is visible in fields of view. The human locations can be represented by the occupancy probability of the ground plane [1]. All cameras are calibrated in advance. Hence, an element of the map at (x, y) corresponds to a rectangle box at one viewpoint. This rectangle means that a human who is 175 cm high at (x, y) will be visible in the rectangle of this view if the person is not occluded by other things. Several methods have been proposed to solve the probabilistic occupancy map (POM). Current methods are mainly based on the concept of “synthesize and compare”. If we guess that persons stand at location i and j , then we can imagine there is a “synthesized” binary map which is union of regions corresponding to location i and j . The goal is to find correct occupied locations such that the “difference” of the background subtraction and the synthesized image is small. In [1], Fleuret et al. tried to minimize the posterior given background subtraction results. The likelihood term is defined by the distance of background subtraction images and synthetic images generated based on occupancy map. An approximated occupancy probability is calculated by minimizing the Kullback-Leibler divergence from true conditional posterior distribution. Alahi et al. [9, 10] constrain the difference of background subtraction results and generated background subtraction results less than a user-defined upper bound of error residual, and l_1 -norm is used to make the solution sparse. These methods only consider background subtraction and neglect any higher-level image information.

In this paper, we develop a novel method fusing these two approaches. First, we modify a general-purpose human detector by using a background image for a camera as the negative training image. These human detectors are used to construct POM. Based on the background subtraction results, we also produce a POM, but we do not generate a synthetic background image. We analyze the relationship between two locations in the ground plane instead. Our method has several advantages. First, these two maps are POMs and their dimensions are much smaller than the size of image. Most previous methods will be also related to the size of image. Our framework is not based on “synthesize and compare” concept; hence, the problem size is just related to the number of interested locations in ground plane, i.e. the dimension of the POM. Secondly, the high-level human detection information is utilized in our method. In addition, the POMs calculated by other methods can be easily incorporated into our framework.

2 Problem Definition and Proposed Method

A ground plane is discretized into several locations, and the probabilistic occupancy map can be arranged as a vector \mathbf{X} . Consider the calibrated camera system with n

views, $\mathbf{I}^1, \mathbf{I}^2, \dots$, and \mathbf{I}^n . We can use the traditional background subtraction method to obtain the foreground map $\mathbf{B}^1, \mathbf{B}^2, \dots$, and \mathbf{B}^n . As illustrated in Figure 1, our method consists of three parts. The first part is to train a scene-aware human detector h_c for each view c , and these c human detectors are used to compute the occupancy probability maps $\mathbf{P}_{\text{HUM}}^c$. The second part is to compute the occupancy probability \mathbf{P}_{BS}^c at a specific location j by calculating the ratio of foreground of the corresponding rectangle \mathbf{R}_j^c in input c -th view, and the relationship of any two locations is considered and summarized in a matrix \mathbf{G}^c . Finally, these two kinds of information, $\mathbf{P}_{\text{HUM}}^c$ and \mathbf{P}_{BS}^c , are integrated into an objective function of \mathbf{X} . The whole problem is cast as a second-order cone programming (SOCP), and many algorithms can solve this type of optimization problems efficiently. The following subsections will introduce these parts.

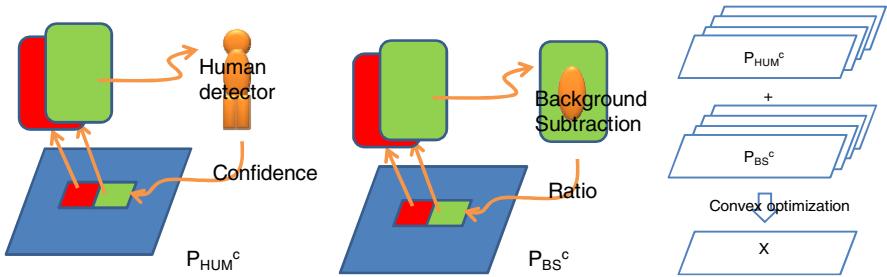


Fig. 1. The overview of our method. The first step is to train a human detector and then create a map $\mathbf{P}_{\text{HUM}}^c$. Secondly, foreground regions are used in generating POM \mathbf{P}_{BS}^c . Finally, these two kinds of information are integrated into an objective function of true occupancy map, and the solution is obtained by convex optimization.

2.1 Scene-Aware Human Detectors

A general-purpose human detector is trained by a set of human images with negative samples which are randomly selected from the public datasets and output the probability of the existence of human in a rectangle. For improving the robustness, we train a human detector h^c by using scene information, called the scene-aware detector. The positive samples are collected from the public database. However, the negative samples are replaced by the images cropped from the background images with no person present in the scene. For each location j on the ground, the corresponding response (the j -th element in $\mathbf{P}_{\text{HUM}}^c$) in the c -th view is calculated by $h^c(\mathbf{R}_j^c)$. Finally, the detector is constructed by AdaBoost-selected EOH features and Haar-like features.

2.2 POM Generation by Background Subtraction

Traditionally, background subtraction is utilized to predict the probability of human occurrence.

We do not use previous complicated methods. A simple construction of POM which captures the information of background subtraction is enough. A human standing at a location i could correspond to a region \mathbf{R}_i^c in view c . We calculate the ratio of foreground to whole region \mathbf{R}_i^c , and we use the value as occupancy probability at location i in view c . If \mathbf{X} is true occupancy, the following term in Equation (1) will not be small, because the regions of neighboring locations will be overlapped.

$$\left\| \mathbf{P}_{BS}^c - \mathbf{X} \right\|_2. \quad (1)$$

We analyze the relationship between two locations. The term in Equation (1) can be modified by a matrix \mathbf{G} . If the two rectangles \mathbf{R}_j^c and \mathbf{R}_k^c are overlapped, then location j will increase the occupancy probability of location k . A matrix, \mathbf{G}^c , can summarize this relationship. The element g_{jk} in the \mathbf{G}^c can be assigned by the ratio of the intersection of \mathbf{R}_j^c and \mathbf{R}_k^c to the rectangle \mathbf{R}_j^c . Hence, a modified term is shown in Equation (2). Originally, \mathbf{X} is a sparse map and it is very different from the map \mathbf{P}_{BS}^c . After modification, the two maps are very similar. Hence, the term can be used in the fusion step.

$$\left\| \mathbf{P}_{BS}^c - \mathbf{G}^c \mathbf{X} \right\|_2. \quad (2)$$

2.3 Fusion of Two POMs

The map \mathbf{P}_{BS}^c are determined based on background subtraction; the map \mathbf{P}_{HUM}^c are calculated from the human detector response $h^c(\mathbf{R}_j^c)$. Solving the true occupancy \mathbf{X} is an inverse problem. These two kinds of maps provide complementary information to determine the true POM. The error of human detectors is defined by

$$\left\| \mathbf{P}_{HUM}^c - \mathbf{X} \right\|_2. \quad (3)$$

The error of background subtraction is more complicated than that of human detectors. The matrix \mathbf{G}^c is incorporated into the error term.

$$\left\| \mathbf{P}_{BS}^c - \mathbf{G}^c \mathbf{X} \right\|_2. \quad (4)$$

By applying the Tikhonov regularization, we have the objective function as follows:

$$\sum_c \left\| \mathbf{P}_{HUM}^c - \mathbf{X} \right\|_2^2 + \gamma \sum_c \left\| \mathbf{P}_{BS}^c - \mathbf{G}^c \mathbf{X} \right\|_2^2 + \delta \|\mathbf{X}\|_2^2, \quad (5)$$

Although this function can be solved easily by the least-squares method, the true occupancy map is very sparse. In order to enforce sparsity, the square of the L2-norm of \mathbf{X} can be replaced by the L1-norm of \mathbf{X} . After replacing square error with the L2-norm, we can use the second-order cone programming (SOCP) formulation for this problem. A general SOCP problem is as follows:

$$\begin{aligned} \min & \quad \mathbf{f}^T \mathbf{x} \\ \text{subject to} & \quad \left\| \mathbf{A}_i \mathbf{x} + \mathbf{b}_i \right\|_2 \leq \mathbf{c}_i^T \mathbf{x} + \mathbf{d}_i, \quad \mathbf{F} \mathbf{x} = \mathbf{g}, \mathbf{x} \geq \mathbf{0} \end{aligned} \quad (6)$$

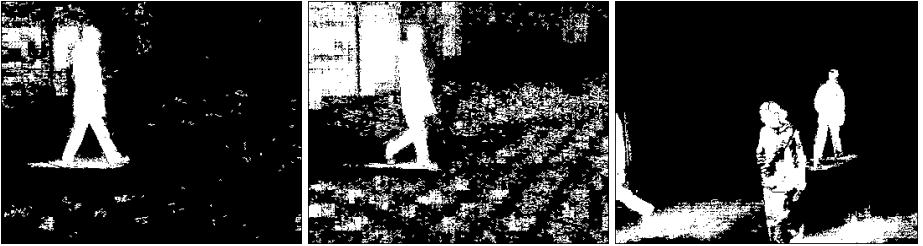


Fig. 2. The degraded background subtraction results for Terrace sequence (the left two images) and Passageway sequence (the right image)

By two slack variables, the associated SOCP problem is given as follows:

$$\begin{aligned} \min \quad & \lambda \|\mathbf{X}\|_1 + \sum_c \alpha_c + \gamma \sum_c \beta_c \\ \text{subject to} \quad & \left\| w^c \left(\mathbf{P}_{HUM}^c \right) - \mathbf{B}^c \mathbf{X} \right\|_2 \leq \alpha_c, \quad \left\| \mathbf{P}_{BS}^c - \mathbf{G}^c \mathbf{X} \right\|_2 \leq \beta_c, \quad \mathbf{X} \geq \mathbf{0} \end{aligned} \quad (7)$$

Because SOCP is a convex optimization problem, the global minimum can be achieved efficiently. In Equation (7), the map can be weighed by a function w^c . This function can be used in increasing contrast (sigmoid function) or identifying importance of the cameras. For example, a camera could have worse quality for distant persons, thus a small weight is assigned to it.

The framework in [9,10] constrain the difference of background subtraction results \mathbf{y} and generated background subtraction \mathbf{Dx} and use l_1 -norm as the objective function. The dictionary \mathbf{D} is a matrix $\mathbf{R}^{M \times N}$, where M is equal to sum of all image pixels and N is the number of locations in ground plane. M is large and proportional to the number of views. In our method, matrices \mathbf{P} , \mathbf{B} , and \mathbf{X} is $\mathbf{R}^{N \times 1}$ and the matrix \mathbf{G} is $\mathbf{R}^{N \times N}$.

$$\begin{aligned} \arg \min_{\mathbf{x}} \quad & \|\mathbf{x}\|_1 \\ \text{subject to} \quad & \|\mathbf{y} - \mathbf{Dx}\|_2 \leq \varepsilon \end{aligned} \quad (8)$$

The methods [1] derived iterative equation to estimate occupancy map, and the methods [9,10] calculated occupancy by minimizing l_1 -norm. Our method use simple POM generated by background subtraction, and main cost in our methods is the fusion step. However, previous methods can also be used in our framework.

3 Experimental Results and Discussion

In surveillance environment, the illumination change, auto white balance and improperly adaptive background mixture models will degrade the results of background subtraction. An example is depicted in Figure 2. If the background subtraction method is reliable, then multiple views with very different settings will be effective to locate people.

Datasets. We test the proposed algorithm on the datasets provided by [1] which is including 6p sequence, Terrace sequence, and Passageway sequence. Each sequence contains four views and ground truth map which is provided by authors.

Measures. We use three measures, precision, recall, and F-measure, in our experiments. Our method will produce a result occupancy map. For comparing our results with ground truth, we have to transform a result occupancy map to a binary map. First, we find the maximum value M and set the threshold T equal to rM , where r is a ratio. In our experiments, we select several ratios, such as 0.99, 0.95, .0.9, ..., etc. All values smaller than rM are set to 0. Then, non-maximum suppression is performed on POM. Finally, we find the maximum matching between result binary map \mathbf{M}_b and ground truth map \mathbf{M}_{gt} . The distance of a matching pair is less than a distance tolerance d_t (typical 1~3, it depends on the density of locations in the ground plane). The precision, recall, F-measure are calculated by the following formulas.

$$\text{precision} = \frac{\# \text{Matching}}{\# \text{Peaks in } \mathbf{M}_b}. \quad (9)$$

$$\text{recall} = \frac{\# \text{Matching}}{\# \text{Peaks in } \mathbf{M}_{gt}}. \quad (10)$$

$$F - \text{measure} = \frac{2(\text{recall})(\text{precision})}{(\text{recall} + \text{precision})}. \quad (11)$$

Background Subtraction and Human Detectors. Our method fuses two POMs which are generated by background subtraction and human detection. The GMM-based [12] background subtraction method [11] with improper parameters (e.g large window size in [11]) is used in our experiments. A human detector can be used to locate people if the camera can capture a whole person or most part of human. In order to compare general purpose and scene-aware human detector, we test these two detectors on subset of Terrace sequence. The performance of these two kinds of detectors is shown in Table 1. The scene-aware detector will have better results. Hence, we use scene-aware detectors in other experiments.

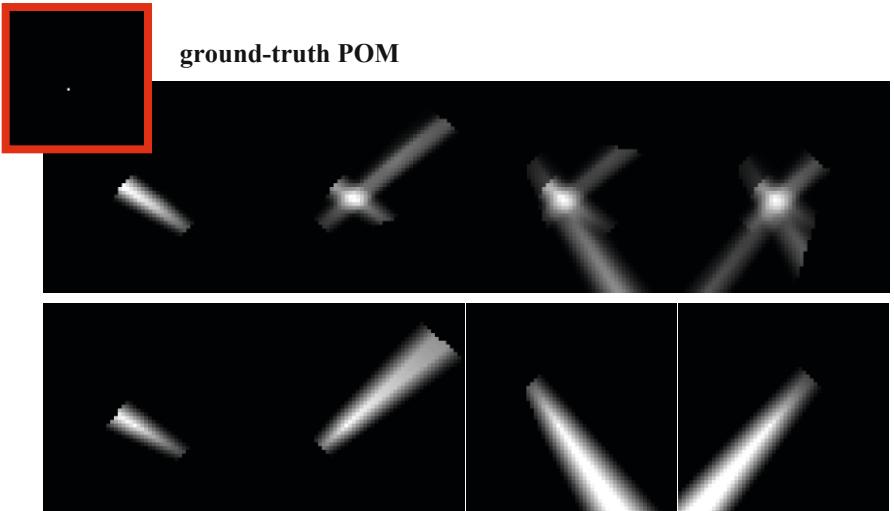


Fig. 3. Top row: the \mathbf{P}_{BS}^c POM generated from the results of background subtraction. Bottom row: multiplication of \mathbf{G}^c and ground truth.

Table 1. Performance for the Human detectors

	F-measure	Precision	Recall
General	0.5873	0.8756	0.4418
Scene	0.5906	0.886	0.4429

Weighting Parameters. In Table 2, we show the results of three settings in the proposed algorithm. BS denotes the setting when the background subtraction is used in the method. It means that the BS method discards α_c and P_{HUM}^c in Equation 5. The HUM setting is to set γ equal to 0. The COMB is the combination of BS and HUM. After minimizing Equation 4, we still need to design a threshold and perform maximum matching. The experimental results show the COMB method will improve F-measure, precision, and recall.

Table 2. Performance of three methods for the Terrace sequence

	F-measure	Precision	Recall
BS	0.4944	0.8109	0.3556
HUM	0.5927	0.8575	0.4528
COMB	0.6182	0.899	0.4711

The results of the other two sequences are shown in Table 3. It seems that our method have worst performance in Passageway1. It is reasonable that dataset 6p is captured in an indoor environment and dataset Passageway1 and Terrace1 are captured in an outdoor environment. Shadow and reflection in Passageway1 will degrade the result of background subtraction.

Table 3. Performance of the COMB for the 6p sequence and Passageway sequence

	F-measure	Precision	Recall
Passageway1	0.4742	0.4012	0.5798
6p	0.6694	0.7050	0.6372

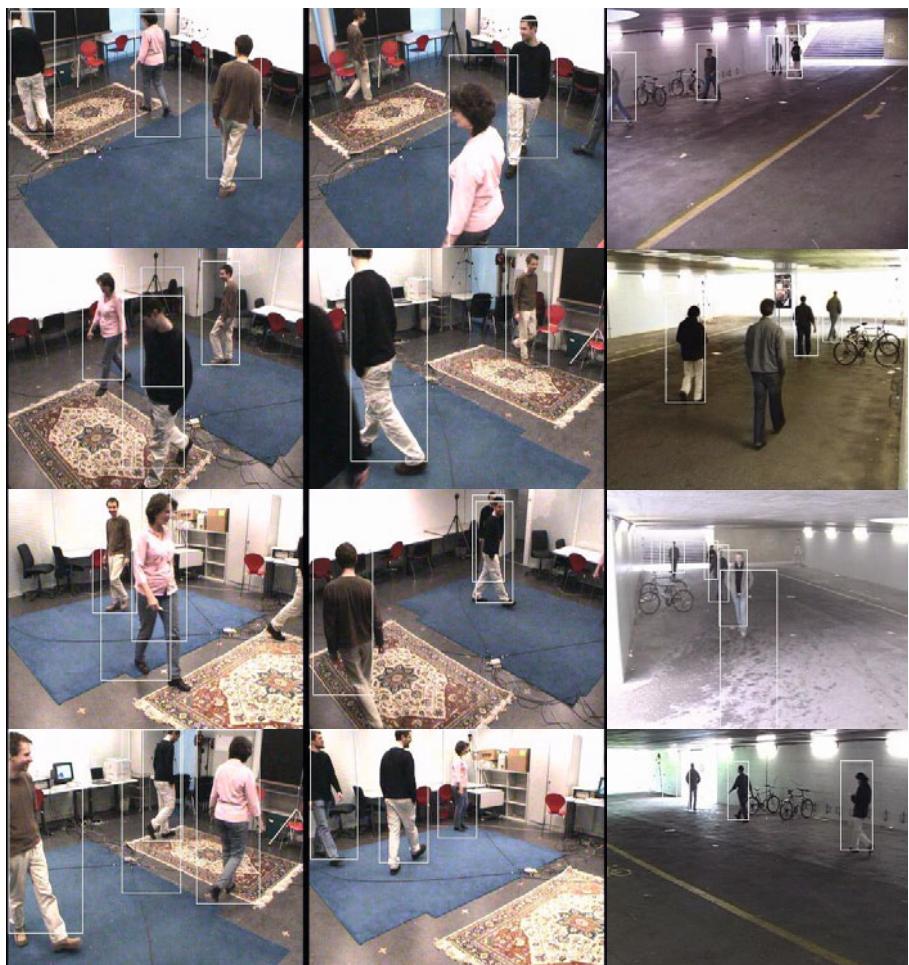
L1-Norm Weighting. In previous experiments, we set λ equal to 0.1. When we set larger λ , the more sparse solution is expected. In Table 4, we tried several λ values for a subset of 6p with more rigorous distance tolerance and we can find that the result has higher recall at $\lambda=0.001$. However, the precision is low at $\lambda=10$. It is because the map is near uniform. Hence, designed threshold could induce too much peaks.

The Matrix G. Here, we will demonstrate that the matrix G is useful for modeling the relationship between locations in the ground. If we do not apply matrix G, the ground truth X is very different from the P_{BS}^c . In Figure 3, we show $G^c X$ and P_{BS}^c . After applying matrix G, $G^c X$ and P_{BS}^c become similar. Finally, we also show some results for Terrace sequences.

Finally, several localization results in 6p and Passageway sequence are shown in Figure 4 and Figure 5. In the first column, there is a false positive. This is because the person in black and the person in pink are close in two views. In the second column,

Table 4. Influence of the weight of $\|l\|_1$ -norm of \mathbf{X} for 6p sequence

Lamda	F-measure	Precision	Recall
0.001	0.4709	0.4460	0.4988
0.01	0.5081	0.4682	0.5555
0.05	0.5258	0.5310	0.5207
1	0.5212	0.5482	0.4967
3	0.4685	0.5547	0.4055
5	0.4234	0.5308	0.3522
7	0.3949	0.5431	0.3103
10	0.3176	0.3561	0.2867

**Fig. 4.** Several localization result for 6p sequence (the first and second columns) and Passage-way sequence (the third column)

all three persons are found out correctly. In Figure 5, the first column is background subtraction result. These two people in two views out of four are close. The second column shows the result of **HUM**. Although there is a false positive in the second column, the result of **COMB** will obtain the correct localization. This example illustrate that combining these two types of complementary information is useful.

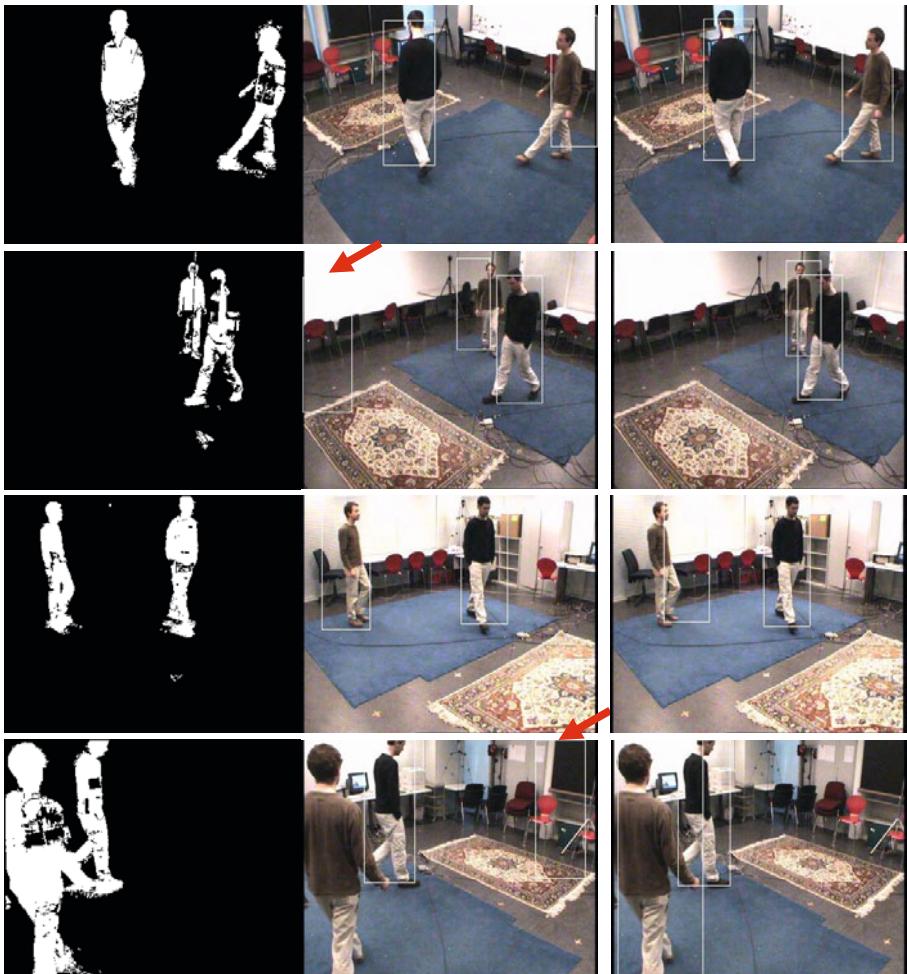


Fig. 5. Several localization result for 6p sequence. The first column is background subtraction result, the second column show the result produce by **HUM**, and the third column is the result of **COMB**.

4 Conclusion and Future Work

We propose a novel method to locate persons under a multi-camera system. Our method uses not only background subtraction but also high-level image information.

We train a scene-aware human detector by using a background image for a camera as the negative training image. The relationship is also analyzed between two locations in the ground plane. Our method has several advantages: these two results are POMs and their dimensions are much smaller than the size of image, most previous methods will be also related to the size of image. Our framework is not based on “synthesize and compare” concept; hence, the problem size is just related to the dimension of the POM. The high-level human detection information is utilized in our method. The experimental results also show the combination of the two complementary POMs works well. The matrix \mathbf{G} and scene-aware human detectors are quite effective.

In this paper, we locate persons from multi-view cameras frame-by-frame. However, the tracking and localization can benefit each other. In the future, we will exploit the relationship between frames and use the information to achieve better human localization.

References

1. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans. Pattern Analysis and Machine Intelligence* 30, 267–282 (2008)
2. Delannay, D., Danhier, N., De Vleeschouwer, C.: Detection and recognition of sports(wo)men from multiple views. In: Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), pp. 1–7 (2009)
3. Berclaz, J., Fleuret, F., Fua, P.: Multi-camera tracking and atypical motion detection with behavioral maps. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part III*. LNCS, vol. 5304, pp. 112–125. Springer, Heidelberg (2008)
4. Levi, K., Weiss, Y.: Learning object detection from a small number of examples: The importance of good features. In: *CVPR*, vol. II, pp. 53–60 (2004)
5. Paisitkriangkrai, S., Shen, C., Zhang, J.: Fast pedestrian detection using a cascade of boosted covariance features. *IEEE Trans. Circuits and Systems for Video Technolog* 18, 1140–1151 (2008)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: *CVPR*, vol. I, pp. 886–893 (2005)
7. Chen, Y., Chen, C.: Fast human detection using a novel boosted cascading structure with meta stages. *IEEE Trans. Image Processing* 17, 1452–1464 (2008)
8. Tuzel, O., Porikli, F., Meer, P.: Pedestrian detection via classification on riemannian manifolds. *IEEE Trans. Pattern Analysis and Machine Intelligence* 30, 1713–1727 (2008)
9. Alahi, A., Boursier, Y., Jacques, L., Vandergheynst, P.: A sparsity constrained inverse problem to locate people in a network of cameras. In: 16th International Conference on Digital Signal Processing, pp. 1–7 (2009)
10. Alahi, A., Boursier, Y., Jacques, L., Vandergheynst, P.: Sport players detection and tracking with a mixed network of planar and omnidirectional cameras. In: Third ACM/IEEE International Conference on Distributed Smart Cameras (ICDSC), pp. 1–8 (2009)
11. Kaewtrakulpong, P., Bowden, R.: An improved adaptive background mixture model for realtime tracking with shadow detection. In: Proc. 2nd European Workshop on Advanced Video Based Surveillance Systems (2001)
12. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. In: *CVPR*, pp. II: 246–252 (1999)

A Novel Depth-Image Based View Synthesis Scheme for Multiview and 3DTV

Xun He, Xin Jin, Minghui Wang, and Satoshi Goto

The Graduate School of IPS, Waseda University
Hibikino, Wakamatsu, Kitakyushu, 808-0135, Japan
goto@waseda.jp

Abstract. Depth-Image Based View Synthesis (DIVS) is considered as a good method to reduce the view numbers for multiview and 3DTV. However, it can't guarantee the quality of occlusion areas of the produced views. In order to reduce the view numbers for multiview and 3DTV, a novel depth-image based view synthesis scheme is proposed to produce four views by one view with depth. Artifact detection and repair functions are added to resolve occlusion areas problem. Artifact region is detected by comparing with original view, and repair function utilizes motion compensation technique to fix it from its previous virtual view. This repair function can correctly fix the artifact region at the cost of some additional bits for each virtual view. In our experiments, only one view is applied to produce four virtual views. About 4~6 dB gain can be achieved at the costs of some additional bit stream. The total bit rate of the four views is only 56% of one color image at the same QP.

Keywords: multiview, 3DTV, view synthesis, depth image.

1 Introduction

In the recent years, there has been a growing interest in 3DTV system, which can enables depth perception for the viewer. Most of available 3-D displays systems just can present one given view point and require wearing specific glasses to perceive the depth of the scene. The future 3-D displays systems are expected to present better stereoscopic view for any view position without glass. One candidate system for future 3D vision is the Philips auto-stereoscopic display [1]. Nine views with eye-distance are required as input and nine views are emitted. Only two views (a stereo pair) can be seen from any position, as in Fig.1. In this system, only one view with depth map is used to produce eight virtual views, which means that occlusions widely exist in the virtual views. Several views and depth maps is needed in order to achieve satisfactory virtual views and to avoid occlusions [11].

Depth-Image Based View Synthesis (DIVS) becomes a key technology for depth based three-dimensional television (3DTV) applications [2]-[3]. Though DIVS can provide a significant improvement in the virtual view synthesis, there are problems of visibility, occlusion, and artifacts. This concept works well for virtual view 5 which is near to the available view such as view 4 and 6. If only one image plus depth view is available, a lot of area will be invisible for the other views. It is not sufficient to fill

invisible holes in these views just by using one reference image. To get better quality virtual view, bi-warping method is proposed, which use two reference views to respectively produce the virtual image and then blend them to generate a final virtual image [4]. The blending effect is better than that in [5]. View Synthesis Reference Software 3.0 (VSRS) also presents a hybrid technique and software for blending based video synthesis system [6] [7]. To improve the quality of virtual views, two views plus depth are used as in Fig.1 (b).

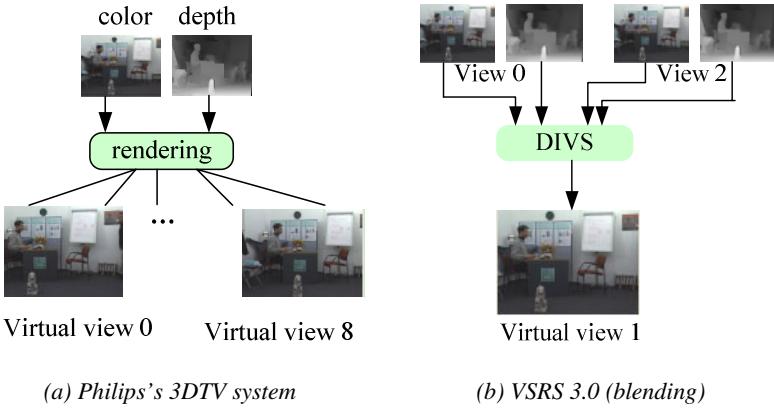


Fig. 1. View Synthesis Scheme

However, multiview system is expected to compress and transmit as few views as possible. These View Synthesis algorithms need two views to produce limited virtual views, which is not quite suitable for reducing bit rate for multiview and 3DTV. In this work, we propose a novel view synthesis scheme, which just needs one view to produce several virtual views. Artifact detection and repair functions are added to resolve invisible region problem. Section 2 briefly describes the novel view scheme. Section 3 is devoted to the details of artifact detection and repairing methods. Section 4 presents the experiment results. And section 5 serves as our conclusion.

2 A Novel View Synthesis Scheme

Efficient compression algorithm in multi-view and 3DTV is an open research issue. The huge amount of data produced by multi-camera systems must be effective compressed in order to reduce the transmitted data. In multiview system, combined temporal/inter-view prediction is applied for multi-view video coding (MVC), where images are not only predicted from temporally neighboring images but also from corresponding images in adjacent views [11]. In 3DTV system, limited views (only one or two views with depth-images) are compressed and transmitted. Decoder reconstructs the other intermediate views by DIVS. Comparing with multi-view system, DIVS based 3DTV needs less bit rate. However, DIVS has some limitation, when it's used for reducing the view number.

In VSRS 3.0, homography matrix [7] is used to transform one 2D view into another. Comparing with projection and re-projection algorithm, it reduces a lot of computational costs. This kind of homography matrix defines 2D transformation of one plane into another one. Then two virtual views are created from the two reference view by using depth map and inverse homography matrix. Two separately synthesized virtual views merge into one, and invisible area can be fixed in the blending processing. There are still some unknown regions in the merged virtual view. They are filled by inpainting algorithm in OpenCV library. VSRS can produce one virtual view from two adjacent views at 37 dB. However, DIVS is very sensitive to the distance between the reference view and target virtual view. There are serious pixel drifting problem in DIVS. As in Fig.2, virtual view 5~7 is produced from view 8. The right part of Fig.2 (a-c) is fetched from original views from position x to $x+49$. The right part is fetched from virtual views from position $x+50$ to $x+87$. Then they are merged together to show the pixel drifting problem. Virtual view 5, 6 and 7 are produce from view 8 in "BookArrival" sequence. A small part of the virtual views are shown together with original view. Fig.2 (a) shows that view 5 has a serious mismatching problem (the character "H" become deformed), as this part of virtual view 5 is about 4 pixels left than original view. Fig.2 (b) shows that virtual view 6 is quite better than virtual view 5, as it's near to the reference view (view 8). It means that we can't produce far distance virtual view with good PSNR.

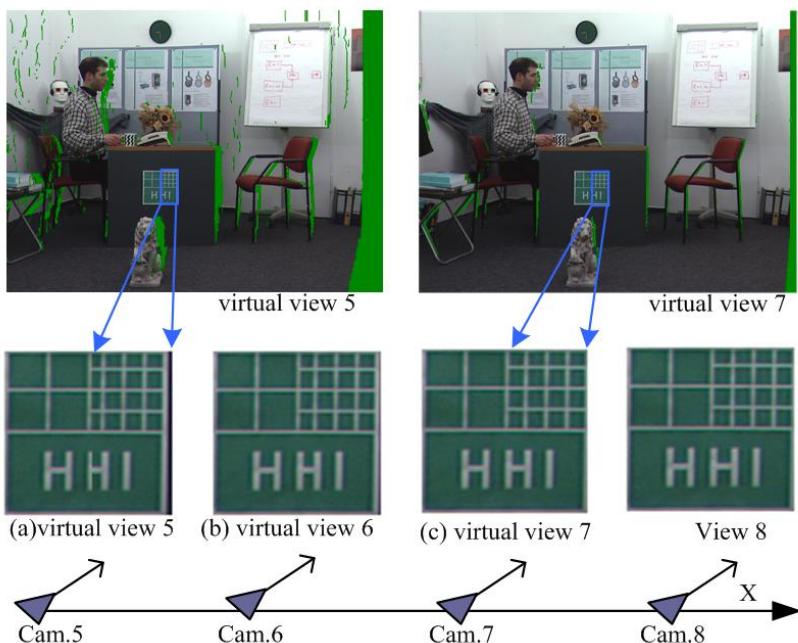


Fig. 2. Pixel drifting and invisible regions in virtual views

Another problem is the invisible areas in the virtual views. In Fig.2, we just use one reference view to produce 3 virtual views. Invisible area and other unsuccessful synthesized area are filled with green color. The invisible area of virtual view 5 is quite larger than virtual view 7, as view 7 is near to view 8. Through we can use inpainting algorithm to fill them, it also causes a lot of artifacts in these area. As this happens at the decoding side, it's hard to predict the correct value for them. However, it's easy to repair them at the encoder side. Once the artifacts area is detected, we can repair it by fetching the corresponding area from the original views.

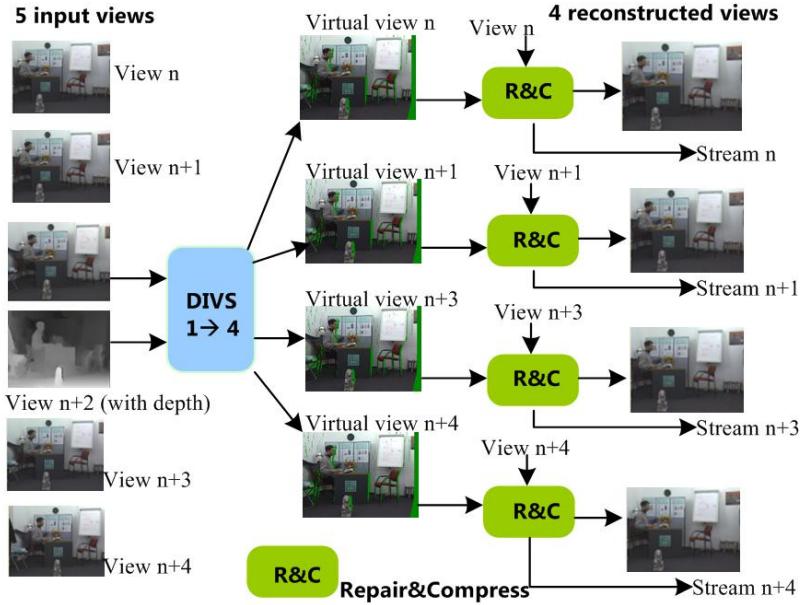


Fig. 3. The proposed view synthesis system

The proposed view synthesis system is shown in Fig.3. View synthesis is performed at both encoding and decoding side. It needs 5 color images and one depth image. The DIVS unit is based on VSRS 3.0 without blending process. R&C unit can detect the artifacts area in the produce virtual view and repair them according to the original view image. The data compression techniques in H.264 are used to compress them. Also there are four additional bit streams for each virtual view. They also should be transmitted together with the bit rate of view n+2. Comparing with conventional view synthesis algorithm, this method transmits fewer views for multiview system. In 3DTV system, this method can provide better quality than previous view synthesis algorithm, as the unreliable or occlusion areas can be fixed.

3 Artifact Detection and Repairing

High quality virtual view without artifacts is critical for the success of view synthesis system. In previous works, view synthesis system is always used at the decoding side

or display system. As there isn't enough information to detect and recover the artifact area. In A. Smolic's proposal for 3DTV [11], they divide the frame into reliable and unreliable regions according to depth boundary. The regions along object boundaries are considered unreliable region. They thought that interpolation artifacts especially occur along object boundaries with depth discontinuities. However, we found that inaccurate depth value also causes a lot of artifacts at smooth depth area as in Fig.4.

The left part in Fig.4 is the synthesized virtual view and the right part in Fig.4 is the original depth value. The blue line is marked at the same position in the two parts. It's obvious that the upper and lower region of the blue line should have different depth value. However, the right part shows that there is no difference in depth image. The depth boundaries based artifact detection algorithm doesn't work in these cases. In our proposed view synthesis system, only one reference view is used for four virtual views. There are a lot of invisible regions which are fixed by inpainting algorithm. In our artifacts detection algorithm, virtual view is first divided into invisible and visible regions.

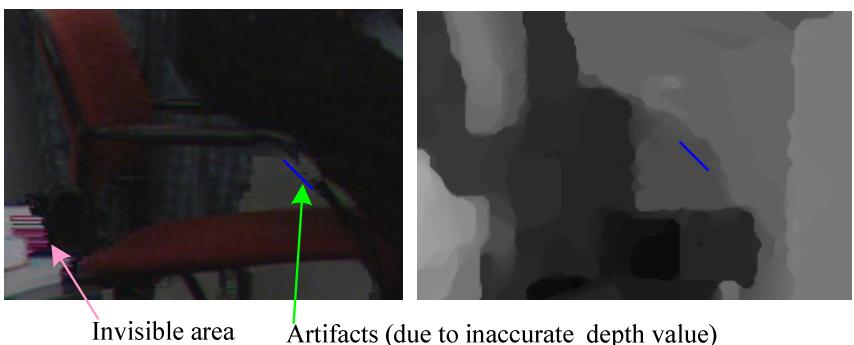


Fig. 4. Artifacts caused by occlusion and inaccurate depth

Invisible region is determined by view synthesis process, as in Fig.2 and Fig.5 (b). The left part of Fig.5 (b) is the original view and the right part shows the invisible region (filled by inpainting). Fig.5 (a) shows the difference between the original view and virtual view in visible region. We can see that difference of visible region is quite large than invisible area, but the subjective quality in visible region is much better than invisible region. Then visible region is always treated as artifacts area.

As the depth image is unreliable as a standard for artifacts detection, color view is used as the most important reference standard to detect artifacts in the visible regions. The artifacts in Fig.4 become easy to detect, as the difference between the virtual view and original view is quite large. In Fig.5 (a), we can see that some background of virtual view is also quite different from the original view (due to sampling error in different camera). It is hard to distinguish them with the real artifacts. Based on this feature we propose a new artifact detection algorithm. The whole frame is divided into 4×4 blocks, called basic processing unit (BPU). As there are pixel drifting problems in virtual view, we define a small search window for checking the difference, a 5×5 area in this work. $Ov(x,y)$ denotes the value at positon (x, y) in original view and

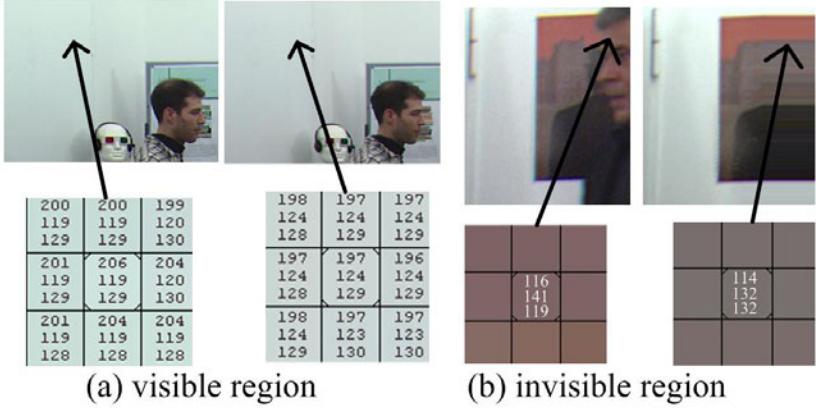


Fig. 5. (a) Difference in visible region (due to sampling error between different cameras); (b) Artifacts in invisible region. Zoom in at the same position.

$Vv(x,y)$ denotes the value at positon (x, y) in virtual view. For BPU at position (m, n) , the minimal DC ($MinDC$) and AC (m, n) ($MinAC$) is defined as in (3) and (4). The $AvgDC(m, n)$ and $AvgAC(m, n)$ in (5) and (6) is the average value of eight neighbouring BPU's $MinDC$ and $MinAC$. In the detection process, the BPU is judged as artifacts when its $MinDC$ or $MinAC$ value is large than Threshold value Ta and Tb as in (7).

$$DC(m,n,\vec{v}) = \sum_{x=m}^{x=m+3} \sum_{y=n}^{y=n+3} Ov(x,y) - \sum_{x=m}^{x=m+3} \sum_{y=n}^{y=n+3} Vv(x,y,\vec{v}) \quad (1)$$

$$AC(m,n,\vec{v}) = \sum_{x=m}^{x=m+3} \sum_{y=n}^{y=n+3} \{Ov(x,y) - Vv(x,y,\vec{v})\}^2 \quad (2)$$

$$MinDC(m,n) = \min(DC(m,n,\vec{v})); (\vec{v}) \in (-2..2) \quad (3)$$

$$MinAC(m,n) = \min(AC(m,n,\vec{v})); (\vec{v}) \in (-2..2) \quad (4)$$

$$AvgDC(m,n) = \sum_{i=m-1}^{i=m+1} \sum_{j=n-1}^{j=n+1} DC(i,j)/8; (i,j) \neq (m,n) \quad (5)$$

$$AvgAC(m,n) = \sum_{i=m-1}^{i=m+1} \sum_{j=n-1}^{j=n+1} AC(i,j)/8; (i,j) \neq (m,n) \quad (6)$$

$$Ta = a \times AvgDC; Tb = a \times AvgAC \quad (7)$$

The whole process is consisted by five steps:

Step (1). Fetch the invisible region information from view synthesis algorithm. Median filter is performed to exclude isolated pixels. Frame is divided into BPUs. If BPU includes isolated pixels, then it's marked as invisible BPU.

Step (2). MinDC and MinAC searching process are performed on visible BPU. This search process just searches the minimal value in a 5x5 region (in original view), as the pixel drifting in virtual view is almost less than 2 pixels.

Step (3). The average value of MinDC and MinAC is calculated as (5) and (6). Two threshold values Ta and Tb are determined by them.

Step (4). When visible BPU's MinDC is larger than Ta or MinAC is larger than Tb, this BPU is marked as artifact BPU.

Step (5). H.264 encoding processing is performed on the MB which has artifacts BPU and invisible BPU. In this work, QP 27 is used for these MBs. Then reconstructed MB replaces this MB.

In our experiment, “a” is equal to 1.3, which is enough to distinguish the background difference as in Fig.5 (a) from the artifacts region. The artifacts region is covered by 4x4 blocks in our proposal. If there are some artifact blocks in a 16x16 macroblock (MB), the MB in original view at this position will be encoded by H.264 encoder. And this MB will be replaced with the reconstructed one. It means that the detected artifacts can be accurately repaired by original view.

If the virtual view synthesis process is just performed at decoding side, detecting and repairing artifacts is quite hard as there is no reference standard. Repairing process always utilizes the neighboring pixels to fill artifact region as in [11], which is quite unreliable. So the most important benefit in this view synthesis scheme is to detect and repair artifacts at encoding side. Robustness can be garneted.

4 Experiments

In our experiments, VSRS 3.0 is used as the basic view synthesis algorithm. View “BookArrival_Cam08” with depth image in “BookArrival” (size: 1024x768) sequence is used to synthesize four neighboring views (view 06, 07, 09 and 10) as in Fig.2. Also the four original views are used in detecting and repairing process. Fig.6 (a) shows part of the produced views before fixing artifacts and Fig.6 (b) shows the repaired views.

To compare the quality with original VSRS, we synthesized view 9 from view 8 and 10. Under this condition, view 9 can achieve about 37dB from two views with depth. Fig.7 shows the PSNR value of these virtual views. Then repaired views can achieve about 4-6 dB gain and achieve similar PSNR with VSRS's best performance. In our experiment, view 9 is encoded by H.264 with QP 27. The bit rate is shown in the right side in Fig.7. The total bit rate of the four views is only 56% of one color image at the same QP.

The proposed method has also been implemented on “Champagne Tower” (1280x960) and “Newspaper”(1024x768) sequences. Table 1 shows the comparisons with VSRS and other view synthesis work [12]. “Champagne”, “Book” and “News” This method can achieve very stable PSNR as it can fix most of obvious artifacts regions by compensation and residual encoding. Comparing with [12], it can achieve 2.9 dB gains (1.2 dB for VSRS) with some extra bit stream. The PSNR performance is quite stable than the others.

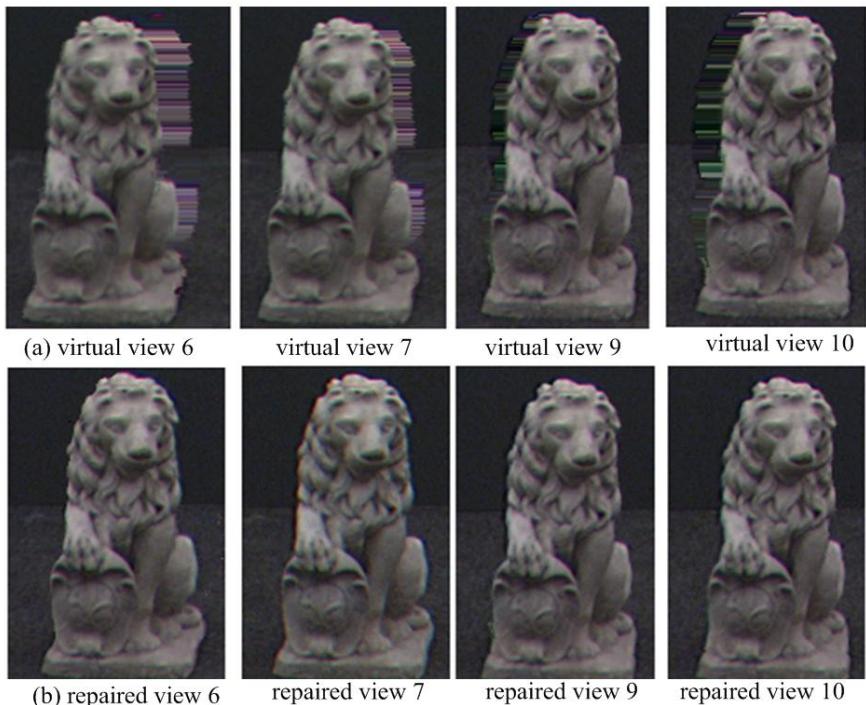


Fig. 6. (a) Virtual views produced from view 08 with depth (invisible region is filled by inpainting); (b) Repaired views by the proposed method according to original view

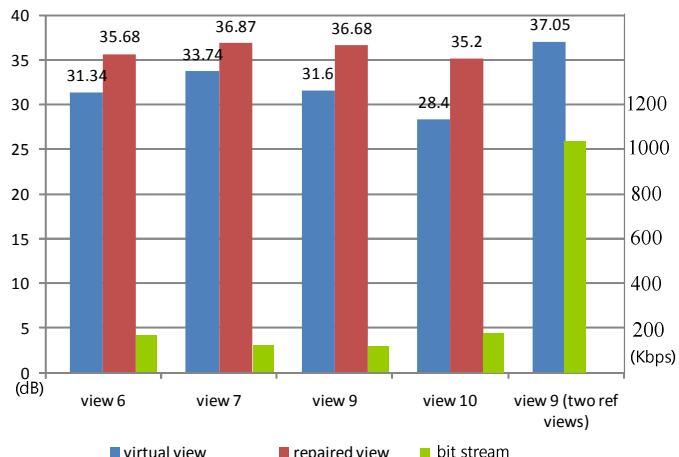


Fig. 7. Quality gain of repaired views and bit costs comparison

Table 1. PSNR (dB) comparisons and bitrate cost of proposal

	Champagne	Book	News
Yang <i>et al</i> [12]	31.92dB	34.67dB	31.87dB
VSRS	32.6dB	37.5dB	33.4dB
Proposed*	35.3dB	36.1dB	35.8dB
bitrate cost	180Kbps	151Kbps	193Kbps

5 Conclusion

Most of previous view synthesis researches just focus on improving quality at decoding or display side. There isn't enough information for detecting and repairing all of the artifacts. However, it's easy in encoding side, as the original views can be used as a reference for detecting and repairing. In experiments we found that long distance view synthesis has pixel drifting problem and encoding the artifacts region only costs a little bit rate. The proposed view synthesis scheme is performed at encoding side, which just needs one view to produce four virtual views. Only the artifacts regions are encoded and transmitted. Four views just needs 56% bit rate of one color view at the same QP.

Acknowledgement

This research was supported by "Ambient SoC Global COE Program of Waseda University" of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. <http://www.philips.com/3Dsolutions>
2. Fehn, C.: A 3D-TV approach using depth-image-based rendering (DIBR). In: Proceedings of Visualization, Imaging, and Image Processing, pp. 482–487 (2003)
3. Fehn, C.: Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV. In: Proceedings of SPIE Stereoscopic Displays and Virtual Reality Systems XI, pp. 93–104 (2004)
4. Liu, Z.-W., An, P., Liu, S.-X., Zhang, Z.-Y.: Arbitrary view generation based on DIBR. In: Proceedings of International Symposium on Intelligent Signal Processing and Communication Systems 2007, pp. 168–171 (2007)
5. Tanimoto, M., Fujii, T., Suzuki, K.: Multi-view depth map of Rena and Akko & Kayo, ISO/IEC JTC1/SC29/WG11, M14888 (October 2007)
6. View Synthesis Reference Software for the 3D Video and Free viewpoint TeleVision project of the 3D Video Coding Team of the ISO/IEC MPEG, version 3.0
7. Video subgroups: View synthesis software and assessment of its performance. ISO/IEC JTC1/SC29/WG11/M15672, Hannover, Germany (July 2008)

8. Merkle, P., Smolic, A., Muller, K., Wiegand, T.: Efficient Compression of Multi-view Depth Data Based on MVC. In: Proc. of IEEE 3DTV Conference (May 2007)
9. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of ACM Conference on Computer Graphics (SIGGRAPH), pp. 417–424 (2000)
10. Shin, H.-C., Kim, Y.-J., Park, H., Park, J.-I.: Fast view synthesis using GPU for 3D display. *IEEE Transactions on Consumer Electronics* 54(2008), 2068–2076
11. Merkle, P., Müller, K., Smolic, A., Wiegand, T.: Efficient Compression of Multi-view Video Exploiting Inter-view Dependencies Based on H.264/MPEG4-AVC. In: IEEE International Conference on Multimedia and Exposition, ICME 2006, Toronto, Ontario, Canada (July 2006)
12. Yang, L., et al.: Probabilistic Reliability Based View Synthesis for FTV. In: IEEE International Conference on Image Processing, ICIP 2010, Hong Kong, September 26-29 (2010)

Egocentric View Transition for Video Monitoring in a Distributed Camera Network

Kuan-Wen Chen¹, Pei-Jyun Lee², and Yi-Ping Hung^{1,2}

¹ Department of Computer Science and Information Engineering

² Graduate Institute of Networking and Multimedia,

National Taiwan University, Taipei, Taiwan

hung@csie.ntu.edu.tw

Abstract. Current visual surveillance system usually includes multiple cameras to monitor the activities of targets over a large area. An important issue for the guard or user using the system is to understand a series of events occurring in the environment, for example to track a target walking across multiple cameras. Opposite to the traditional systems switching the camera view from one to another directly, we propose a novel approach to egocentric view transition, which synthesizes the virtual views during the period of switching cameras, to ease the mental effort for users to understand the events. An important property of our system is that it can be applied to the situations of where the view fields of transition cameras are not close enough or even exclusive. Such situations have never been taken into consideration in the state-of-the-art view transition techniques, to our best knowledge.

1 Introduction

Multi-camera systems have been widely used in video surveillance applications, such as airport or railway security, traffic monitoring, and etc. The main benefit of multi-camera system is that it can monitor the activities of targets over a large area. However, to security guards or users using the system, with the number of video streams increasing the difficulty of monitoring increases, especially when an event happening among multiple cameras. For example, a common surveillance activity is to track a target that moves from the field of views (FOVs) of one camera to another.

It is a common practice to show multiple video streams on display simultaneously, like a screen wall. However, the characteristics of human vision system allow human to pay attention to one video at a time [15]. When tracking a target moving to another camera view, such an approach requires the user to “physically” adjust the view point to the display window of where the target appears. In addition, the user needs to “mentally” translate [5][22] the view point from one camera to another to understand where the target is in the environment and the geometrical relationship between cameras. Hsiao et al. [5] hypothesized that human produces the pictorial-based representation of mental image [3] when switching views. This mental transition will be very difficult if the user is new to the environment, the environment is sophisticated, or even the camera network is large and widely distributed.

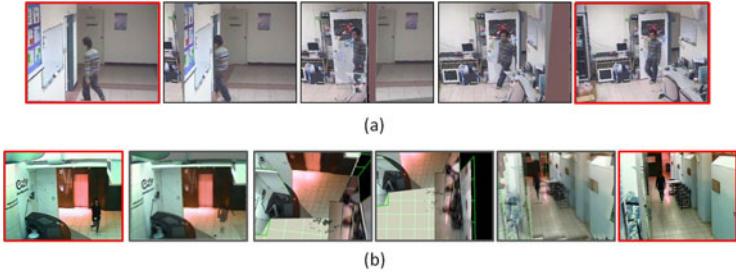


Fig. 1. An example of egocentric view transition between (a) overlapping cameras and (b) non-overlapping cameras. The left-most image and the right-most image are the images captured by real cameras, and the middle images are the synthesized transition images from the camera in the left-most column to the camera in the right-most column.

To ease intensive mental effort for users, in this paper, we propose a novel approach to egocentric view transition, which synthesizes the virtual views between cameras, as shown in Fig. 1. Users can always focus on the target even when switching cameras, and it avoids the effect of uncomfortable flash caused by sudden view change, a traditional way to switch camera views. Furthermore, it helps users understand the spatial relationships among the target, cameras, and environments easily.

Our method deals with both overlapping and non-overlapping cameras. In the case with overlapping cameras, opposite to the existing view transition techniques, our method does not need to match feature correspondences, and thus can work real-time and can transit camera views even when cameras are in wide base-line situations. In the case with non-overlapping cameras, we visualize the probability distribution of where the target is in the blind region with a particle-based approach.

In the following section, when view transition from Camera 1 to Camera 2, the Camera 1 and Camera 2 are denoted as *start-camera* and *end-camera* for simplicity, respectively.

2 Related Works

Little attention has been paid to help humans monitor the events across multiple cameras. The basic concept of view transition comes from view morphing [18], and there have been many applications of the morphing techniques, such as virtual teleconference system [12], sports broadcasting system [10], and photo browsing and exploring system [21]. However, these methods are usually applied to the views whose FOVs are overlapped each other and feature correspondences are needed to be estimated. Thus, they cannot work in real time and are difficult to be applied to visual surveillance applications with the cameras usually installed distant to each other.

To monitor multiple cameras, some works embedded video surveillance images in a 3D model by using projective texture mapping [20] to integrate live video streams with the model, for example augmented virtual environment system [13], video flashlight technique [17], and DOTS system [4]. Such works produced some unnatural effects, when an object was projected to different parts of the model. Wang et al. [24] pasted the

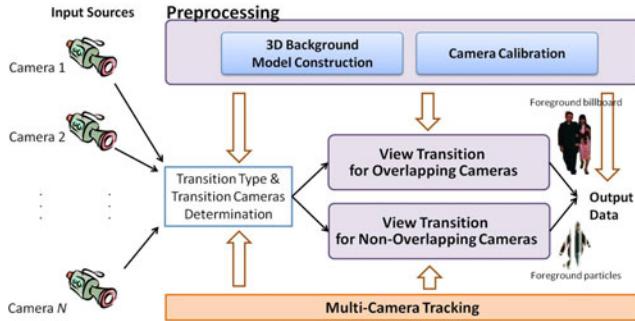


Fig. 2. System overview of our approach to egocentric view transition

image of each camera to a plane and put the plane at the model. The methods mentioned above need to construct a complete 3D background model in advance. In addition, this kind of visualization is with a gap for users between the view of cameras and the view in the 3D model or map, and so users still have to mentally translate the focusing view when tracking a target across cameras [22].

To avoid such mental effort, the egocentric view transition techniques are developed. Katkere et al. [9] firstly proposed a framework to synthesize the virtual views when switching cameras in a surveillance system. Ohta et al. [14] proposed a billboarding technique to transit views in a soccer stadium. Recently, Haan et al. [7] transited views with dynamic video embedding method, which rendered video images in a 3D environment, similar to Wang's method [24], and allowed smooth transition even in a complex environment.

The system proposed by Haan et al. [7] is most similar to what we present in this paper. However, their approach does not deal with the foreground object, and hence the foreground target will disappear during the period of switching views. This target's disappearance will cause a discontinuous effect to users, especially when the FOVs of two cameras are not close enough or even exclusive.

3 System Overview

As shown in Fig. 2, our system includes three main technical components: preprocessing, view transition for overlapping cameras, and view transition for non-overlapping cameras.

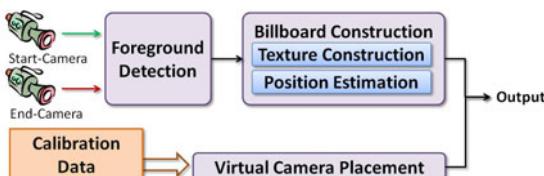


Fig. 3. The flowchart of view transition between overlapping cameras

When system starts up and a transition event happens, the system determines the transition type is an overlapping case or a non-overlapping case and the related transition cameras. This process can be taken easily from the information of the preprocessing component. Then, our system runs either “view transition for overlapping cameras” part or “view transition for non-overlapping cameras” part. A brief description of the view transition parts is to produce foreground texture, put it onto the model, and determine the position of virtual camera. Finally, the output image is generated by blending the synthesized foreground texture into the pre-constructed background model and then re-projecting it to the view of virtual camera.

3.1 Preprocessing

The preprocessing component is an offline process and contains two parts. The first part is “background 3D model construction.” With the concept of view dependent texture mapping [2], only a simple background 3D model is required, including the walls and ground planes. In our implementation, we construct the model manually, which can also be estimated by using multi-view geometry [8]. After constructing the environment model, we map the background texture image of each calibrated camera onto the model with projective texture mapping technique [20].

The second part is “camera calibration.” The intrinsic parameters of each camera are obtained by Zhang’s method [25]. The relative orientation and pose to the constructed model are calculated by the nonlinear refinement with the Levenberg-Marquardt Algorithm [11] given several correspondences of 3D points and 2D points.

3.2 Multi-camera Tracking

In our system, we track targets across cameras and switch cameras automatically by applying a multi-camera tracking algorithm. In the overlapping case, multi-camera tracking is performed by comparing the 3D positions estimated from each camera. When the positions of two targets are close enough, we consider that they are the same target. In the non-overlapping case, we implement the method proposed by Chen et al. [1], which tracks targets across non-overlapping cameras based on both spatio-temporal and appearance cues.

4 View Transition for Overlapping Cameras

The view transition for overlapping cameras happens when a tracked target is within the FOVs of both transition cameras, and the other camera has a better view to monitor the target. The flowchart is shown in Fig. 3. The input data includes two camera images and the camera parameters. The output data includes a billboard with foreground texture, the position of billboard in 3D model, and the virtual camera position.

4.1 Foreground Detection

The foreground detection component is to extract the foreground pixels in the image of each camera. To detect foreground, a pixel-level background modeling algorithm [19] is used, which models the background pixels as a Gaussian Mixture Model

(GMM) and the weights of the mixture and the parameters of the Gaussians are adapted with respect to the current frames. After that, some post-processing techniques are applied. The shadow pixels are detected and removed by Horprasert's method [6]. The morphological operation is used to remove some isolated noise. Finally, we use a smooth operation to avoid the effect of obvious edges after blending the background model and foreground texture.

4.2 Foreground Billboard Construction and Position Estimation

After the foreground detection, a foreground alpha mask and its corresponding texture can be extracted. According to the alpha mask value, the foreground texture is pasted to a billboard. Then, we blend the billboards extracted from each real camera into a final one with different blending weights, which is inversely proportional to the distance between the position of real cameras and that of virtual camera.

To estimate the 3D position of foreground target, we make an assumption here. We suppose the target is standing on the ground plane of the environment. With this assumption, we can compute the 3D position of foreground target by using only one camera. First, we estimate the foot position in 2D image of camera from the foreground detection result. Second, we project the 2D point to a 3D point on the ground plane of model with camera parameters known. Furthermore, to avoid irregular movement of the foreground object caused by the noise of foreground detection, Kalman filter [23] is used to smooth the estimated trajectory finally.

4.3 Virtual Camera Placement

In the overlapping case, the virtual camera position can be determined easily. In most situations, the virtual camera parameters are linearly interpolated from the parameters of start-camera and end-camera, and the transition can perform well. The only exception is when two real cameras are looking in almost opposite directions. If the linear interpolation method is also used, the virtual camera will pass through the region close to the target and cause a strange effect. Thus, we let the virtual camera navigate along an arc trajectory with centering on the target in such situation.

5 View Transition for Non-overlapping Cameras

The view transition among non-overlapping cameras starts when a tracked target is leaving the FOV of the focusing camera (start-camera) and entering a blind region. Then we set the virtual camera to monitor the blind region and visualize the probability distribution of where the target is with a particle system. Particle system [16] is widely used in the field of computer graphics and applied to model fuzzy objects such as water and smoke. When a target appears in one camera's view, our system determines whether the target is the same as that disappearing before. If they are the same target, the camera will be set as end-camera and the focusing view will switch from virtual camera to it automatically.

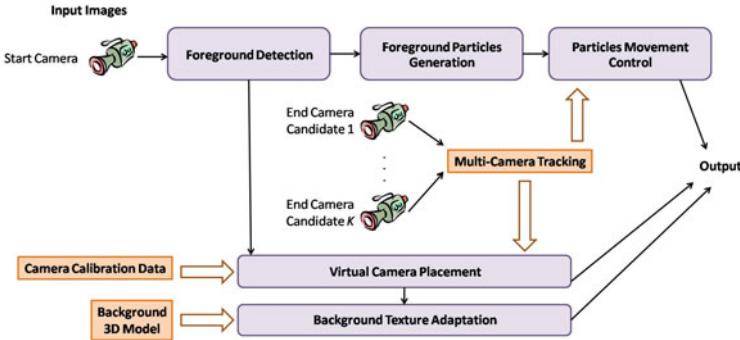


Fig. 4. The flowchart of view transition among non-overlapping cameras

Fig. 4 shows the flowchart of view transition among non-overlapping cameras. The input data comes from one start-camera and multiple end-camera candidates. The end-camera candidates are the cameras whose FOVs are next to the blind region. The output data includes the textures and positions of foreground particles, the virtual camera position, and the texture of background model. In the non-overlapping case, the texture of background model is adapted to the position of virtual camera to avoid some model walls with coarse texture or even no texture mapped from the real camera images.

5.1 Foreground Particles Generation

This component describes how to generate new particles in our system. We model each particle as a long thin rectangle with foreground texture on it. After foreground detection, we divide the foreground texture region into several long thin rectangles. We assign the long thin texture to each particle and generate the particle in the position of where the foreground texture corresponds to, as shown in Fig. 5(c).

5.2 Particles Movement Control

There are three modes of particle movement in our implementation: initial movement, diffuse movement, and shrink movement. In the initial movement mode, the particles move slowly in random directions to let the particles disperse uniformly and are duplicated during the period.

In the diffuse movement mode, the particles diffuse in the blind region. To simulate the probability distribution of where the target is, we assign each particle a velocity according to the velocity distribution. The velocity distribution is calculated by dividing the distance by the corresponding transition time. The distance can be obtained from the pre-constructed model. The transition time probability distribution is learnt by Chen's method [1].

After the multi-camera tracking module detects the target appearing in one camera, the mode of particle movement is changed to the shrink movement mode. During the period, the number of particles is decreasing, and the particle will be given a large velocity to shrink to the place of target in the view of end-camera quickly.

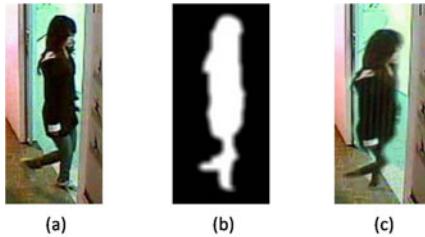


Fig. 5. The process of foreground particle generation: (a) original image, (b) foreground detection result, and (c) the generated foreground particles on the background

5.3 Virtual Camera Placement

To view transition among non-overlapping cameras, an important issue is how to set the virtual camera position to let the user monitor the blind region without feeling uncomfortable. Here, we propose a rule to set three main positions of virtual camera. It avoids the virtual camera passing through the wall, lets the user always look at the diffuse region of particles, and avoids the view angle of camera changing too fast. When view transition, the virtual camera will move along a linearly interpolated path through these three main positions sequentially, as shown in Fig. 6.

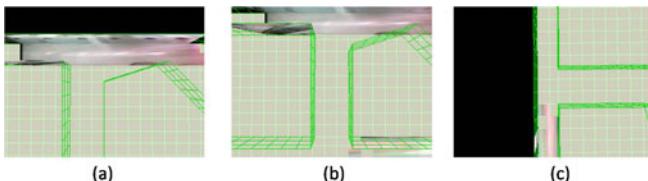


Fig. 6. An example of three main virtual camera positions (transiting views from camera 3 to camera 8 in Fig. 8(b)): (a) the first position, (b) the second position, and (c) the third position

The first position is the upper place of where the target entering the blind region with camera looking downward and the optical line vertical to the ground plane. The second position is the upper place of the center of position distribution of particles. The third position is the upper place of where the target leaving the blind region.

5.4 Background Texture Adaptation

In the non-overlapping case, some parts of model are not captured by real cameras. Moreover, some texture of model comes from cameras would be very coarse, because the distance between the camera and the texture region may be very large, as shown in Fig. 7(a). Therefore, we develop an adaptive grid-based visualization to represent the regions with coarse texture or no texture. It substitutes grid-texture for real texture, when the texture of a region is with less pixel density in the image of real camera than that captured by virtual cameras.

We calculate the pixel density of texture ratio of real camera to virtual camera by the following equation:

$$R_r = \frac{A_{RC}}{A_{VC}}, \quad (1)$$

where A_{RC} and A_{VC} are the areas of a region r in 3D model projecting to the images of real camera and virtual camera, respectively. When R_r is larger than 1, the texture of the region r is pasted with that captured by cameras. Otherwise, the grid-texture is used, as shown in Fig. 7(c).

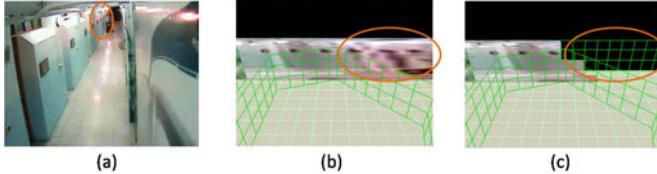


Fig. 7. (a) The original image, (b) the image of virtual camera without grid-based visualization, (c) the image of virtual camera with grid-based visualization. The orange ellipse represents the corresponding region in three images.

6 Results

There are two environments in our experiments. The first environment is shown in Fig. 8(a). There are four cameras and the FOVs of cameras are overlapping. Although only overlapping case is considered in this environment, there are three different view transition scenario here: (1) a long passage transition (transition from camera 1 to camera 2), (2) a corner transition (transition from camera 2 to camera 3), and (3) an indoor-outdoor transition (transition from camera 3 to camera 4). Notice that the views of two cameras are quite different in the second and third scenario. Therefore, those previous works (e.g. [18][12]) that need to find feature correspondences are no

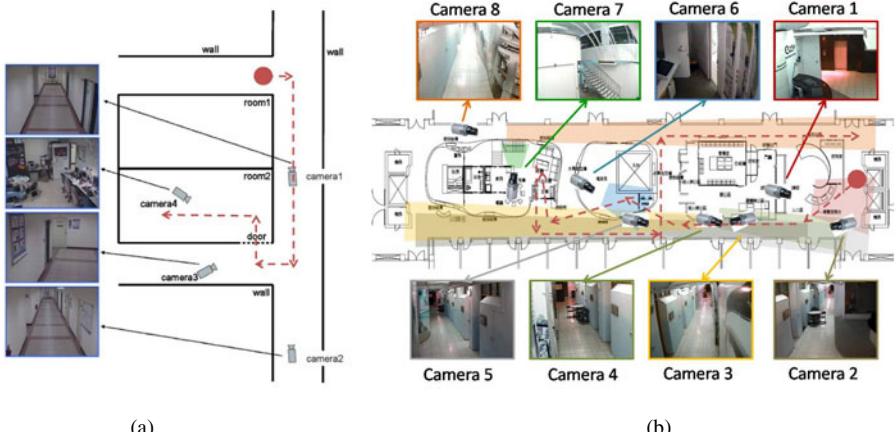


Fig. 8. (a) The first experimental environment with four cameras. (b) The second experimental environment with eight cameras and their corresponding FOVs. The red dotted line is the path of target in our testing video sequence.

longer viable. The second environment (Fig. 8(b)) includes eight cameras. Some of them are with overlapping FOVs, but there are still many blind regions in the monitored environment.

Some of the results are shown in Fig. 9 and Fig. 10. The left-most column and the right-most column are real images, and the other middle columns are synthesized view transition images from the camera in the left-most column to the camera in the right-most column. More details can be seen in the video sequence “<http://www.csie.ntu.edu.tw/~f93014/EgocentricViewTransition.wmv>.”

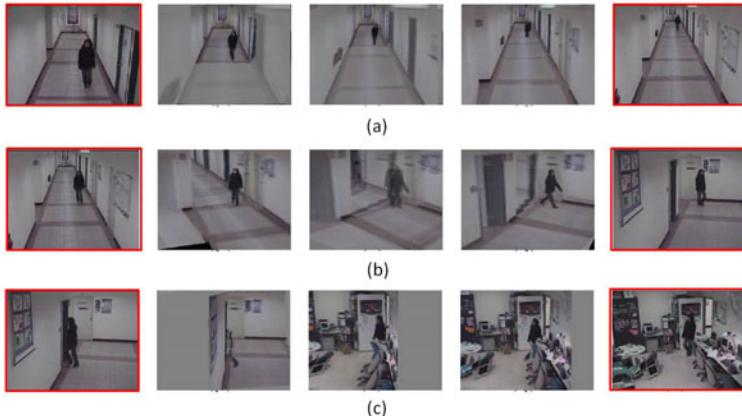


Fig. 9. The view transition results in the first environment: (a) transition from camera 1 to camera 2, (b) transition from camera 2 to camera 3, and (c) transition from camera 3 to camera 4

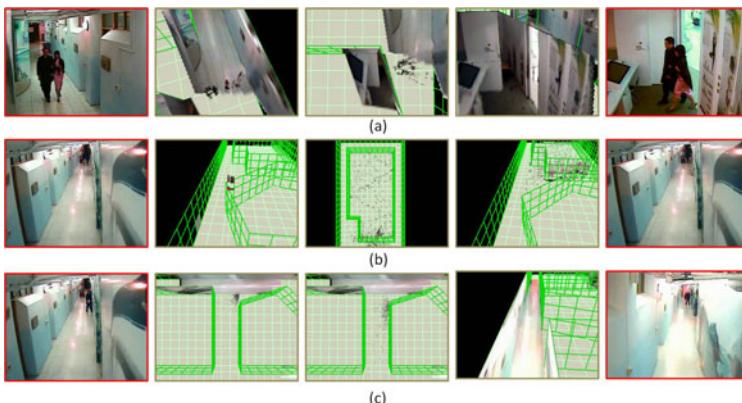


Fig. 10. The view transition results in the second environment: (a) transition from camera 5 to camera 6, (b) transition from camera 3 to a blind region and then back to camera 3, and (c) transition from camera 3 to camera 8

7 Conclusions

We have proposed an approach to egocentric view transition which synthesizes the virtual views when switching cameras. Our method dealt with the environments

including both overlapping and non-overlapping FOVs of cameras. In the case of view transition between overlapping cameras, we presented a framework to build a foreground billboard and put it to the 3D model. In the case of view transition among non-overlapping cameras, we proposed a framework to use a particle system to visualize the probability distribution of where the target is in the blind region. In addition, a rule of setting virtual camera positions and a background texture adaptation method were developed for a better view transition effect. In the experiments, we installed our system in two environments. It demonstrated that our method can perform well in both overlapping and non-overlapping cases.

Although our current work can be applied to most environments well, there are still some issues needed to be solved in the future. First, a better way to determine the path of virtual camera can be explored by doing some user tests. Second, a user study is needed to demonstrate the foreground billboard and particles are indeed helpful for user to track the target across cameras. Third, with better computer vision techniques, the effect of foreground texture can be improved, especially in a complex environment.

Acknowledgements

This work was supported in part by the Ministry of Economic Affairs, Taiwan, under Grant 99-EC-17-A-02-S1-032, and the Excellent Research Projects of National Taiwan University, under grants 99R80303.

References

- Chen, K.W., Lai, C.C., Hung, Y.P., Chen, C.S.: An Adaptive Learning Method for Target Tracking across Multiple Cameras. In: CVPR (2008)
- Debevec, P.E., Taylor, C.J., Malik, J.: Modeling and Rendering Architecture from Photographs: A Hybrid Geometry- and Image-Based Approach. In: SIGGRAPH (1996)
- Finke, R.A.: Principles of Mental Imagery. MIT Press, Cambridge (1989)
- Girgensohn, A., Kimber, D., Vaughan, J., Yang, T., Shipman, F., Turner, T., Rieffel, E., Wilcox, L., Chen, F., Dunnigan, T.: DOTS: Support for Effective Video Surveillance. In: MULTIMEDIA (2007)
- Hsiao, C.H., Huang, W.C., Chen, K.W., Chang, L.W., Hung, Y.P.: Generating Pictorial-Based Representation of Mental Image for Video Monitoring. In: IUI (2009)
- Horprasert, T., Harwood, D., Davi, L.: A Statistical Approach for Real-Time Robust Background Subtraction and Shadow Detection. In: FRAME-RATE Workshop (1999)
- Haan, G., Scheuer, J., Vries, R., Post, F.H.: Egocentric Navigation for Video Surveillance in 3D Virtual Environments. In: IEEE Symposium on 3D User Interfaces (2009)
- Hartley, R.I., Zisserman, A.: Multiple View Geometry, 2nd edn. Cambridge University Press, Cambridge (2004)
- Katkere, A., Moezzi, S., Kuramura, D.Y., Kelly, P., Jain, R.: Towards Video-Based Immersive Environments. Multimedia System 5(2), 69–85 (1997)
- Kanade, T., Narayanan, P., and Rander, P.: Virtualized Reality: Concept and Early Results. Tech. Rep. CMU-CS-95-153 (1995)
- Levenberg, K.: A method for the solution of certain problems in least squares. Quarterly Applied Math. 2, 164–168 (1944)

12. Lei, B., Hendriks, E.: Real-Time Multi-Step View Reconstruction for a Virtual Teleconference System. *EURASIP J. Appl. Signal Process* 2002(10), 1067–1088 (2002)
13. Neumann, U., You, S., Hu, J., Jiang, B., Lee, J.: Augmented Virtual Environment (AVE): Dynamic Fusion of Imagery and 3d Models. In: *IEEE Virtual Reality* (2003)
14. Ohta, Y., Kitahara, I., Kameda, Y., Ishikawa, H., Koyama, T.: Live 3D Video in Soccer Stadium. *IJCV* 75(1), 173–187 (2007)
15. Palmer, S.: *Vision Science: Photons to Phenomenology*. MIT Press, Cambridge (1999)
16. Reeves, W.T.: Particle Systems - a Technique for Modeling a Class of Fuzzy Objects. *ACM Transactions on Graphics* 2, 91–108 (1983)
17. Sawhney, H.S., Arpa, A., Kumar, R., Samarasekera, S., Aggarwal, M., Hsu, S., Nister, D., Hanna, K.: Video Flashlights: Read Time Rendering of Multiple Videos for Immersive Model Visualization. In: *EGRW* (2002)
18. Seitz, S., Dyer, C.: View Morphing. In: *SIGGRAPH* (1996)
19. Stauffer, C., Grimson, W.E.L.: Learning Patterns of Activity using Real-Time Tracking. *IEEE Transactions on PAMI* 22(8), 747–757 (2000)
20. Segal, M., Korobkin, C., Widenfelt, R., Foran, J., Haeberli, P.: Fast Shadows and Lighting Effects Using Texture Mapping. In: *SIGGRAPH* (1992)
21. Snavely, N., Seitz, S.M., Szeliski, R.: Photo Tourism: Exploring Photo Collections in 3d. In: *SIGGRAPH* (2006)
22. Thorndyke, P., Hayes-Roth, B.: Differences in Spatial Knowledge Acquired from Maps and Navigation. *Cognitive Psychology* 14(4), 560–589 (1982)
23. Welch, G., Bishop, G.: An introduction to the kalman filter. Chapel Hill, NC, USA, Tech. Rep. (1995)
24. Wang, Y., Krum, D.M., Coelho, E.M., Bowman, D.A.: Contextualized Videos: Combining Videos with Environment Models to Support Situational Understanding. *IEEE TVCG* 13(6), 1568–1575 (2007)
25. Zhang, Z.: A Flexible New Technique for Camera Calibration. *IEEE Transactions on PAMI* 22, 1330–1334 (2000)

A Multiple Camera System with Real-Time Volume Reconstruction for Articulated Skeleton Pose Tracking

Zheng Zhang¹, Hock Soon Seah¹, Chee Kwang Quah²,
Alex Ong³, and Khalid Jabbar³

¹ School of Computer Engineering, Nanyang Technological University, Singapore
`zhang_zheng@pmail.ntu.edu.sg`, `ashsseah@ntu.edu.sg`

² Institute for Media Innovation, Nanyang Technological University, Singapore
`quah_chee_kwang@rp.sg`

³ Republic Polytechnics, School of Sports, Health and Leisure, Singapore
`{alex_ong,khalid_jabbar}@rp.sg`

Abstract. We present a multi-camera system for recovering skeleton body pose, by performing real-time volume reconstruction and using a hierarchical stochastic pose search algorithm. Different from many multi-camera systems that require a few connected workstations, our system only uses a single PC to control 8 cameras for synchronous image acquisition. Silhouettes of the 8 cameras are extracted via a color-based background subtraction algorithm, and set as input to the 3D volume reconstruction. Our system can perform real-time volume reconstruction rendered in point clouds, voxels as well as voxels with texturing. The full-body skeleton pose (29-D vector) is then recovered by fitting an articulated body model to the volume sequences. The pose estimation is performed in a hierarchical manner, by using a particle swarm optimization (PSO) based search strategy combined with soft constraints. 3D distance transform (DT) is used for reducing the computing time of objective evaluations.

Keywords: Multi-Camera System, Real-time Volume Reconstruction, Pose Estimation.

1 Introduction

Marker-based motion capture has been used in a variety of applications. One important application is in films, animation or video games, where it is used for recording and mapping the movements of actors to animate characters. Another application is for sport analysis and biomechanics, where motion capture provides cost-effective solutions in the application of rehabilitation, patient positioning, or sport performance enhancement. Motion capture has also been applied to surveillance systems, for example, to analyse actions, activities, and behaviors both for crowds and individuals. In recent years, due to the emerging techniques and research in computer vision and computer graphics, a variety of markerless approaches to motion capture have been proposed [4,26,2,17,10,16]. These

markerless approaches trying to recover 3D body pose from images without using markers or special suits show potential to replace the traditional marker-based systems though many issues still remain unresolved.

Approaches of markerless human motion capture can be classified by the number of cameras they use, i.e., monocular or multi-view approach. Compared to monocular approaches, multi-view approaches have the advantage that they can deal better with occlusion and appearance ambiguity problems, leading to much more robust and accurate results. These approaches can be model-based or not. The model-based methods, which use a priori model of the subject to guide the pose tracking and estimation processes, can greatly simplify the problem and make the pose estimation more robust and accurate. In this research project, we are interested in recovering human skeleton motion from multi-view image sequences using a model-based method.

One framework of a model-based multi-view approach is to fit a predefined model to the images by aligning the projection of the model with some sets of 2D image features, such as silhouettes [2], contours [12], color or texture. Another framework is to directly perform pose estimation and tracking in the 3D space, by using 3D reconstruction data such as stereo [29] or volume [4,17,10]. Working in 3D space has several advantages. First, it brings more consistency, while tracking in 2D domain is more easily affected by self-occlusions. Second, 3D data such as volume synthesizes all the information concerning the camera parameters and background subtraction, allowing simpler and more efficient tracking.

In this paper, a synchronized multi-camera system (Sect. 3) is described that uses only one PC to control 8 cameras for capturing multiple image streams. Silhouettes of the 8 cameras are extracted via a color-based background subtraction algorithm, and set as input to the 3D volume reconstruction. Our system can perform real-time volume reconstruction (Sect. 4). The full-body skeleton pose (29-D vector) is then recovered by fitting an articulated body model to the volume sequences via a PSO based hierarchical pose search algorithm (Sect. 5).

2 Previous Work

Multi-camera systems that can provide synchronized multi-view video streams has been widely used for image-based rendering [2], scene reconstruction [21], and motion analysis [22,6]. Such a system usually relies on a few connected PCs or workstations.

In recent years, markerless motion capture with multi-camera system has been paid much attention. One line of capturing motion from multi-camera images is to reconstruct the time-series volumes of the moving person by using shape-from-silhouette or visual hull [14] methods, and then acquire the internal motion from the volume sequences [4,3,25,17,10,16]. In the early works, Cheung et al. [4], C. Theobalt et al [25], and Mikic et al. [17] present systems that can track human body using voxels with acceptable robustness. In the recent works, many systems rely on more sophisticated pose search startegies for reovering body pose from volume sequences. Kehl et al. [10] present to use the stochastic meta descent algorithm (SMD) incorporated both volume and color information for pose

tracking. Their later work [11] combines more other image cues such as edges, color information together with volume information. Mundermann et al. [18] employ an ICP algorithm to fit a chosen SCAPE body model against visual hull sequences. Similarly, Ogawara et al. [19] create a deformable mesh-based model from visual hull method and match this model with the volume data through the ICP algorithm with Kd-tree search in pose and normal space. Both [1] and [8] use a Bayesian tracking framework and show real-time pose recovery from volumetric sequences. However, their methods rely on using a learned motion behavior model to bias the propagation of particles that limits its applicability.

Our work is most close to the work of [10] where a descent (SMD) optimization is used to fit an articulated body model to the volume sequences reconstructed from multi-view streams. Our approach uses a global stochastic (PSO) optimization algorithm in a hierarchical pose search framework, combined with soft space constraints and 3D DT map for increasing the robustness and efficiency.

3 The Multi-camera System

3.1 System Setup

Building a multiple camera system is not an easy task as it involves great effort on the hardware as well as the software development. Many considerations should be taken into account [25]: First, all cameras must be able to work synchronously for acquiring multiple image sequences that are correctly registered in time. Second, the frame rate of image acquisition should be at least 15fps; otherwise, the system may not be applied to scenes when the subject performs fairly fast movements. Third, the bandwidth should be sufficient for supporting the transfer of multi-video streams. Finally, the acquisition room ought to be large enough to allow multi-view recording of dynamics scenes from a sufficient distance and from a large number of viewpoints.

Considering the requirements, we set up our multiple camera system in an acquisition room of appropriate size, with static background and stable illumination. This indoor environment provides good condition for silhouette extraction.

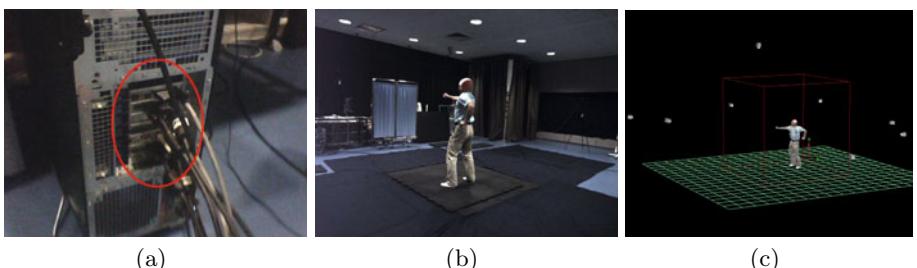


Fig. 1. (a) Single PC with 4 PCI cards. (b) Capture room with one subject performing movements inside the interest space. (c) Camera configuration: 8 cameras are installed with different point views.

We use eight Point Gray cameras including three Flea 2 and five Grasshopper IEEE1394b cameras. These CCD cameras are triggered to synchronously capture RGB color images at a resolution of 800×600 pixels at 15 fps. Different from many multi-camera systems that require many connected workstations (e.g., [25][13][24]), our multi-camera system only uses a single PC to control 8 cameras (Fig. 1(a)). We use 4 PCI cards including two 2-port IEEE-1394b PCI express cards which support 160 MB/s data transfer, and two 4-port IEEE 1394b FireWire 800 PCI cards which supports 400MB/s data transfer. The high speed and wide bus bandwidth makes the system capable of providing synchronized image sequences from all the 8 cameras in a single PC, and avoiding the legacy of low network transmission occurring on multi-PCs systems. Fig. 1 gives an illustration of the workstation, capture room and camera configuration.

3.2 Camera Calibration

Camera calibration is performed as follows: we first calibrate the intrinsic camera parameters using small chessboard images ((Fig. 2(a))) via the widely used Zhang's algorithm [28], then compute the extrinsic parameters from the image of a large chess grid (Fig. 2(b)) which is placed on the floor area that can be viewed by all cameras from different view points. The world coordinate system is set at a corner of the grid, with xy plane being the floor.

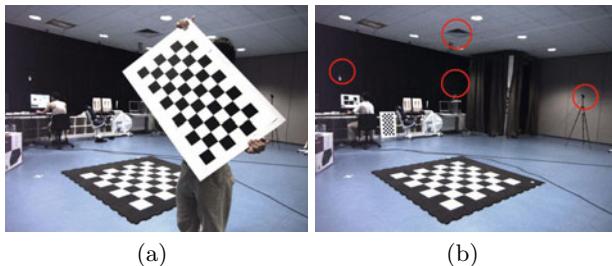


Fig. 2. (a) Chess board for intrinsic parameter calibration. (b) Chess grid for extrinsic parameter calibration (four cameras are circled out).

4 Volume Reconstruction

4.1 Background Subtraction

To extract the silhouettes, we use a background subtraction approach that identifies moving foreground objects from the background. The indoor capture environment has global or local illumination changes such as shadows and highlights. Shadows cast from foreground subject onto the environment are easily incorrectly classified as foreground. We adopt a RGB-color based background subtraction method [7] which can efficiently remove the shadow and highlight.

4.2 Shape-from-Silhouette and Visual Hulls

Shape-from-silhouette is a method of obtaining the visual hull which is the upperbound to the actual volume of the subject [14]. Several approaches have been proposed to compute visual hulls. They can be roughly separated into two categories: polyhedral and volumetric approach. Polyhedral approaches generate polyhedral visual hulls [15]. In this case, a 2D polygon approximation of each multi-view silhouette contour is firstly obtained, and then these silhouette polygons are back-projected into the space from their corresponding camera positions. The final object's visual hull is computed as the intersection of the silhouette cones. While in the case of volumetric approaches [23], the 3D space is divided into cubic volume elements (i.e., voxels) and a test is conducted for each voxel to decide whether the voxel lies inside or outside the object's visual hull. The test is performed by checking the overlap of each voxel's projections with all the multi-view silhouettes. The final reconstruction result is a volumetric representation of the object.

For our motion capture application, we use a volumetric approach to obtain the volume data. Volumetric approach is more robust since they do not rely on the silhouette contours but on the occupied/non-occupied image regions. Additionally, its reconstruction can be easily limited to the desired accuracy. One problem is the high computational cost. Fortunately, the use of dedicated hardware and software acceleration techniques makes real-time volume reconstruction possible [13,4].

We use a similar volume reconstruction method as that of Cheung et al. [4]. In contrast with their method, we only project the center rather than the eight corners of each voxel to the silhouette for the intersection test. The projection of the center point of each voxel is pre-computed and stored in a lookup table for computational speeding up. The method will output the reconstruction result that contains both inside and surface voxels of the whole volume of the subject. As the surface voxels are enough for the pose estimation application, a post-process is needed to remove the inside voxels. This can be done effectively by testing each voxel's 6-connected neighbors. If the type of any one of the neighbors

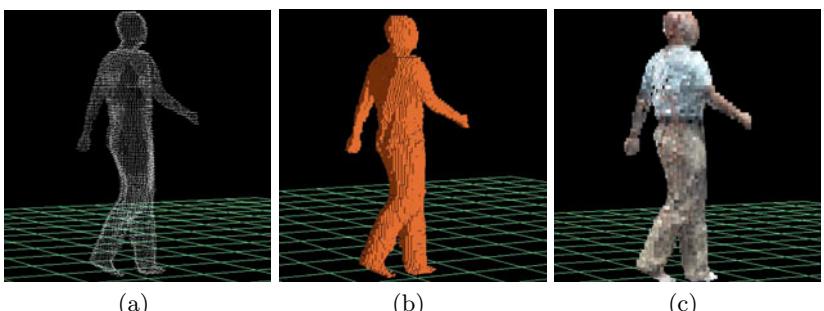


Fig. 3. Illustration of volume reconstruction rendered in point clouds (a), voxels without texturing (b) and voxels with texturing (c)

Table 1. Algorithm of Volume Reconstruction

-
1. Initialization: subdivide the 3D space of interest into $N \times N \times N$ equal sized cubes and compute the center point coordinate for each cube. Then build a lookup table for storing the pre-computed projection data.
 2. For each camera view, perform a thorough scan:
 - (1) For each cue:
 - (2) If the type of the cube is “0”, continue; Otherwise conduct the point-silhouette-intersection test.
 - (3) If the projection of the center point is inside the silhouette, set the cube’s type as “1”; otherwise, set as “0”.
 - (4) End for.
 3. Remove the inside voxels.
-

is 0, the voxel is a surface voxel. This post-process can save much CPU time in rendering. The main steps of the algorithm are listed in Table 1.

In our system, the textures of the multi-camera views can be optionally incorporated to the voxel data, by mixing all colors of the views where the voxel is not occluded. Fig. 3 illustrates one sample of volume reconstruction result rendered in point clouds, voxels and voxels with texturing.

5 Skeleton Pose Estimation

5.1 The Body Model

A parametric shape primitive named barrel model is used to represent body parts which are connected by joints in kinematic chains to describe the articulated body structure. This body model consisting of 10 body segments, each of which has a local coordinate system to define the locations of the points that are regularly sampled from surface points of the barrel shape. The body segments are connected by joints to form five open kinematic chains with a total of 29 DOFs.

5.2 Pose Estimation and Tracking

The system performs pose estimation in a hierarchical manner, with space constraints integrated into a PSO pose search algorithm. PSO is a stochastic searching technique of population-based evolutionary computation, and has been successfully used for pose recovery [9, 27] and many hard numerical and combinatorial optimization problems [20, 5]. It relies on the below two equations:

$$\mathbf{v}_{k+1}^i = \chi(\mathbf{v}_k^i + \mathbf{U}(0, \phi_1) \otimes (\mathbf{p}_i - \mathbf{x}_k^i) + \mathbf{U}(0, \phi_2) \otimes (\mathbf{p}_g - \mathbf{x}_k^i)) \quad (1)$$

$$\mathbf{x}_{k+1}^i = \mathbf{x}_k^i + \mathbf{v}_{k+1}^i \quad (2)$$

where \mathbf{x}_k^i and \mathbf{v}_k^i separately denote the position and velocity of the i -th particle at k -th iteration, χ is a constriction coefficient, $\mathbf{U}(0, \phi_i)$ represents a vector of

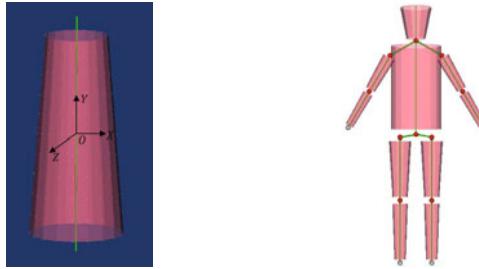


Fig. 4. Illustration of the articulated body model (R) built from barrel models (L)

random numbers uniformly distributed in $[0, \phi_i]$, \otimes denotes a point-wise vector multiplication. \mathbf{p}_i is the history best position found by the i_{th} particle and the \mathbf{p}_g is the global best position found by its neighborhood so far.

The hierarchical pose search is performed as a sequence of sub-optimizations on the body parts. We first compute the six DOFs of the torso. The three rotations of the head joint are then estimated using the same optimization method. After that, we turn to independently fit the four limbs each of which has five DOFs. Each sub-optimization process is performed by fitting the corresponding body model part using the volume cue. For example, if the sub-optimization is to compute the pose parameters of the root joint, the fitting should be performed on the torso of the body model. Or if the sub-optimization is for the joints of a leg, then the thigh and calf of the leg will be taken into account.

To obtain robust results, the temporal information is incorporated into the objective function:

$$E(\mathbf{x}) = w_V E_V(\mathbf{x}) + w_T E_T(\mathbf{x}) \quad (3)$$

where w_V and w_T are two weight factors used to balance the influence of the two objective errors. The first term E_V pushes the body model to match the volume data. The second term E_T improves the temporal smoothness of the motion

$$E_T(\mathbf{x}) = \|\mathbf{x} - \dot{\mathbf{x}}\|^2 \quad (4)$$

where $\dot{\mathbf{x}}$ is the pose of the previous frame. To reduce the computational cost, 3D DT map is used for the closest correspondence finding and closest distance computing.

Two space constraints are taken into account when computing the objective error $E_V(\mathbf{x})$. First, different body segments are not allowed to intersect in the space. Second, different model points should avoid taking the same closest feature point. The two constraints are integrated into the objective function by introducing the penalty factor λ . If the model point \mathbf{p} lies in a voxel bin that has been occupied by other model points from a different body segment, the value of λ is set as λ_1 ; or else if the model point \mathbf{p} has a closest feature voxel that has been taken by other points, λ is set as λ_2 ; otherwise it is set as 1. Both λ_1 and λ_2 are larger than 1; their actual values can be determined by experiments.

For the first frame, pose initialization problem should be addressed. Our approach solves this by first interactively obtaining the global position and then

using a similar search algorithm described above that without the term temporal smoothing. For pose tracking, the swarm particles are propagated according to a weak motion model (Gaussian distribution), which enhances particle diversity for keeping the search ability of the swarm.

$$\boldsymbol{x}_{t+1}^i \sim \mathcal{N}(\boldsymbol{p}_t^i, \boldsymbol{\Sigma}) \quad (5)$$

where \boldsymbol{p}_t^i is the best position of i -th particle at time step t . We assume $\boldsymbol{\Sigma}$ is a diagonal covariance matrix such that $\boldsymbol{\Sigma} = \mathbf{I}\delta_j^2$. The standard deviations δ_j are related to the frame rate of the captured image sequences as well as the body parts' motion. For simplicity, we set δ_j as a constant value δ and derive it by experiments.

6 Results

The system is tested on several sequences of one subject performing in the volume space. The volume reconstruction with color texture rendering runs at about

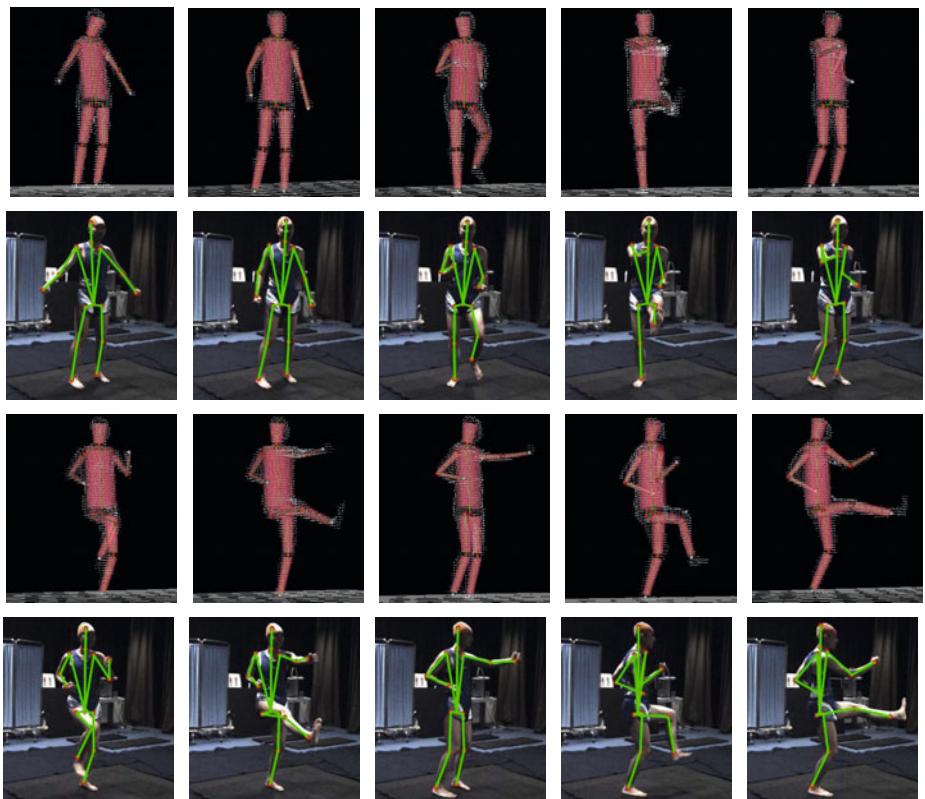


Fig. 5. Pose tracking example results of a sequence of one subject

5 fps for $100 \times 100 \times 100$ voxel resolution. For lower resolution, for example, $64 \times 64 \times 64$, it can reach 7.5 fps.

For pose tracking, the PSO parameter χ is set as 0.7298, ϕ_1 and ϕ_2 are set as 2.05 according to the Clerc's constriction PSO algorithm [5]. The maximum iteration for each swarm is set to be the same as the pose initialization, except for the 6-DOFs sub-optimization of torso which is set as 80. The penalty factors λ_1 and λ_2 are separately set as 20 and 5 which are optimally determined by experiments. The objective function (Equ. 3) depends on two weight factors w_V and w_T . We experimented with random changes to the two parameters and use $w_V = 1$ and $w_T = 10$ for our sequences that are recorded at 15 fps. For sequences that have faster movements or low frame rate, w_T should be set to a smaller value. We set δ for swarm propagation (Equ. 5) as 15 and limit the maximum joint angle standard deviations to 40 degrees. All volume sequences for motion tracking are reconstructed at a voxel resolution of $80 \times 80 \times 80$. Fig. 5 show some pose tracking results of one subject.

7 Conclusion

In this work, we build a synchronized multi-camera system that relies on single PC to control 8 IEEE1394b cameras. Real-time volume sequences are reconstructed for articulated pose recovery. To track full body movements over several frames, our approach fits an articulated body model to the volume data, using a PSO-based hierarchical search algorithm combined with soft space constraints. We remark that the tracker would meet failures for the volume frames where the body limbs are close to the torso. This is a common problem for volume based method [3]. To obtain good tracking performances, the articulated model should be well suited the subject's body shape. Future work will concentrate on enhancing the tracking robustness and accurateness.

References

1. Caillette, F., Galata, A., Howard, T.: Real-time 3-d human body tracking using variable length markov models. In: Proceedings of British Machine Vision Conference, BMVC (2005)
2. Carranza, J., Theobalt, C., Magnor, M.A., Seidel, H.-P.: Free-viewpoint video of human actors. In: ACM SIGGRAPH 2003 (2003)
3. Cheung, G.K.M., Baker, S., Kanade, T.: Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture. In: IEEE Conference on Computer Vision and Pattern Recognition (2003)
4. Cheung, G.K.M., Kanade, T., Bouguet, J.-Y., Holler, M.: A real time system for robust 3d voxel reconstruction of human motions. In: IEEE Conference on Computer Vision and Pattern Recognition (2000)
5. Clerc, M., Kennedy, J.: The particle swarm - explosion, stability, and convergence in a multidimensional complex space. IEEE Transaction on Evolutionary Computation 6, 58–73 (2002)

6. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.-P., Thrun, S.: Performance capture from sparse multi-view video. In: ACM SIGGRAPH 2008 papers, SIGGRAPH 2008, pp. 1–10. ACM, New York (2008)
7. Horprasert, T., Harwood, D., Davis, L.S.: A statistical approach for real-time robust background subtraction and shadow detection. In: IEEE ICCV 1999, pp. 1–19 (1999)
8. Hou, S., Galata, A., Caillette, F., Thacker, N., Bromiley, P.: Real-time body tracking using a gaussian process latent variable model. In: IEEE International Conference on Computer Vision (2007)
9. Ivezkovic, S., Trucco, E.: Human body pose estimation with pso. In: IEEE Congress on Evolutionary Computation (2006)
10. Kehl, R., Bray, M., Van Gool, L.: Full body tracking from multiple views using stochastic sampling. In: IEEE Conference on Computer Vision and Pattern Recognition (2005)
11. Kehl, R., Van Gool, L.: Markerless tracking of complex human motions from multiple views. *Computer Vision and Image Understanding* 104(2), 190–209 (2006)
12. Knossow, D., Ronfard, R., Horaud, R., Devernay, F.: Tracking with the kinematics of extremal contours. In: Asian Conference on Computer Vision, pp. 664–673 (2006)
13. Ladikos, A., Benhimane, S., Navab, N.: Efficient visual hull computation for real-time 3d reconstruction using cuda. In: Computer Vision and Pattern Recognition Workshop, pp. 1–8 (2008)
14. Laurentini, A.: The visual hull concept for silhouette-based image understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* 16(2), 150–162 (1994)
15. Matusik, W., Buehler, C., McMillan, L.: Polyhedral visual hulls for real-time rendering. In: Proceedings of the 12th Eurographics Workshop on Rendering Techniques, pp. 115–126 (2001)
16. Michoud, B., Guillou, E., Briceno, H., Bouakaz, S.: Real-time marker-free motion capture from multiple cameras. In: IEEE International Conference on Computer Vision (2007)
17. Mikic, I., Trivedi, M., Hunter, E., Cosman, P.: Human body model acquisition and tracking using voxel data. *International Journal of Computer Vision (IJCV)* 53, 199–223 (2003)
18. Mundermann, L., Corazza, S., Andriacchi, T.P.: Accurately measuring human movement using articulated icp with soft-joint constraints and a repository of articulated models. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
19. Ogawara, K., Li, X., Ikeuchi, K.: Marker-less human motion estimation using articulated deformable model. In: IEEE International Conference on Robotics and Automation (2007)
20. Parsopoulos, K.E., Vrahatis, M.N.: Recent approaches to global optimization problems through particle swarm optimization. *Natural Computing* 1, 235–306 (2002)
21. Starck, J., Hilton, A.: Model-based multiple view reconstruction of people. In: IEEE International Conference on Computer Vision (2003)
22. Starck, J., Hilton, A.: Surface capture for performance-based animation. *IEEE Computer Graphics and Applications* 27(3), 21–31 (2007)
23. Szeliski, R.: Rapid octree construction from image sequences. *CVGIP: Image Understanding* 58(1), 23–32 (1993)
24. Takahashi, K., Nagasawa, Y., Hashimoto, M.: Remarks on 3d human body's feature extraction from voxel reconstruction of human body posture. In: IEEE International Conference on Robotics and Biomimetics (2007)

25. Theobalt, C., Li, M., Magnor, M., Seidel, H.-P.: A flexible and versatile studio for synchronized multi-view video recording. In: Vision, Video, and Graphics (2003)
26. Theobalt, C., Magnor, M., Schüler, P., Seidel, H.-P.: Combining 2d feature tracking and volume reconstruction for online video-based human motion capture. In: Proceedings of the 10th Pacific Conference on Computer Graphics and Applications, p. 96 (2002)
27. Ivezović, Š., Trucco, E., Petillot, Y.R.: Human body pose estimation with particle swarm optimisation. *Evolutionary Computation* 16(4), 509–528 (2008)
28. Zhang, Z.: A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 1330–1334 (1998)
29. Ziegler, J., Nickel, K., Stiefelhagen, R.: Tracking of the articulated upper body on multi-view stereo image sequences. In: *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 774–781 (2006)

A New Two-Omni-Camera System with a Console Table for Versatile 3D Vision Applications and Its Automatic Adaptation to Imprecise Camera Setups

Shen-En Shih¹ and Wen-Hsiang Tsai^{1,2}

¹ Institute of Computer Science and Engineering, National Chiao Tung University, Taiwan
peter159.cs98g@nctu.edu.tw

² Department of Information Communication, Asia University, Taiwan
whtsai@cis.nctu.edu.tw

Abstract. A new two-omni-camera system for 3D vision applications and a method for adaptation of the system to imprecise camera setups are proposed in this study. First, an efficient scheme for calibration of several omni-camera parameters using a set of analytic formulas is proposed. Also proposed is a technique to adapt the system to imprecise camera configuration setups for in-field 3D feature point data computation. The adaptation is accomplished by the use of a line feature of the console table boundary. Finally, analytic formulas for computing 3D feature point data after adaptation are derived. Good experimental results are shown to prove the feasibility and correctness of the proposed method.

Keywords: omni-directional camera, camera calibration, adaptation to imprecise camera setups, 3D data computation, computer vision applications.

1 Introduction

With the advance of technologies, various types of omni-directional cameras have been used for many applications, like virtual and augmented reality, TV games, video surveillance, 3D environment modeling, etc. In order to interact with human beings, most applications require acquisition of 3D data of feature points. This usually means in turn the need of a precise camera setup to yield accurate 3D data computation results. However, from the viewpoint of a non-technical user, it is not reasonable to ask him/her to set up the cameras at right places and orient them to right directions without errors. Therefore, the system should be designed to have a capability to adapt to camera setup errors *automatically*.

In this study, we first propose a novel method to calibrate omni-cameras, which is more efficient than traditional methods. Next, we propose a new 3D vision platform for various applications, as shown in Fig. 1, which consists of two omni-cameras, a console table, and a display (see Fig. 1(a) for an illustration). Each camera is affixed to the top of a rod, forming an “omni-camera stand.” The two stands are placed beside the console table, which is located in front of the display (see Fig. 1(b)). The display may be a TV, a large monitor, a projection screen, etc. The height of the rod of each

omni-camera stand is adjustable to fit the height of the console table and/or that of the user of the system. While placing the omni-camera stands beside the table, the cameras are supposed to be oriented exactly forward to face the user and used to take omni-images of the user's activity in front of the table (see Fig. 1(c)). The two cameras are assumed to be well calibrated *in the factory* by the proposed calibration method, and the optical axes of the two omni-cameras are supposed to be mutually parallel and both horizontal with respect to the floor. If the omni-camera stands are set up precisely in this way *in the field*, then the 3D data computation can be conducted precisely as well. As an example of applications of the proposed system for 3D data computation, Fig. 1(c) shows the case of a user using a finger tip of his, covered with a yellow cot, as a *3D cursor point*, which is useful for 3D space exploration in video games, virtual or augmented reality, 3D graphic designs, and so on. Fig. 1(d) shows the extracted finger tip region and Fig. 1(e) shows the centroid point of the region whose 3D location is to be computed in this study.

However, it is assumed in this study that the system, after being manufactured in factory, is made available to a common in-field user who does not have the knowledge of camera calibration. He/she is allowed to put the camera stands beside the console table freely. So the optical axes of the two cameras may not be mutually parallel. It is desired in this study that the system can be designed to be *adaptable* to such a situation *automatically* by using a line feature of the console table boundary as a landmark to compute the angle between the two optical axes. In this way, the system can still conduct 3D data computation in terms of the computed angle for various applications. This is a new topic which has not been studied so far according to our survey of the literature.

Many 3D vision applications using multiple omni-directional cameras have been exploited, such as human tracking [1], 3D reconstruction [2], moving object localization [3], etc. In addition, traditional calibration methods use planar calibration patterns and extract point features from them to calculate the focal length of the camera [7]. In the proposed method, we calculate the focal length in terms of some parameters of the omni-camera structure which may be measured conveniently. There are also some works of calibrating an omni-camera using features on the omni-camera itself [4-6]. And a method to detect a space line in an omni-image, which is adopted for use in this study, was proposed in [8].

In summary, the proposed method has the following merits: (1) being able to calibrate the parameters of the omni-camera efficiently; (2) having the capability to detect automatically relevant parameters under imprecise system setups and adapt the system function accordingly; and (3) being capable of conducting 3D data computation after system adaptation.

In the following, we first describe the idea of the proposed method in Section 2, and present the details of the method in Sections 3 through 5. Some experimental results are shown in Section 6, followed by conclusions in Section 7.

2 Idea of Proposed Method

The proposed method includes three main stages: (1) in-factory camera calibration; (2) in-field system adaptation; and (3) 3D data computation. The first stage is conducted in an in-factory environment with the goal of calibrating the camera parameters

efficiently. For this, a novel method is to be proposed using some conveniently measurable system features. The second stage is conducted in an in-field environment where the optical axes of the two omni-cameras are not adjusted to be mutually parallel. In this stage, an in-field calibration process will be done automatically by using a line feature of the console table boundary. With that line feature, the proposed method can calculate the angle between the two optical axes. The computed angle is then used in computation of the 3D data of a feature point on the user's hand in the third stage, achieving the adaptation of the system function for 3D data computation under an imprecise camera setup.

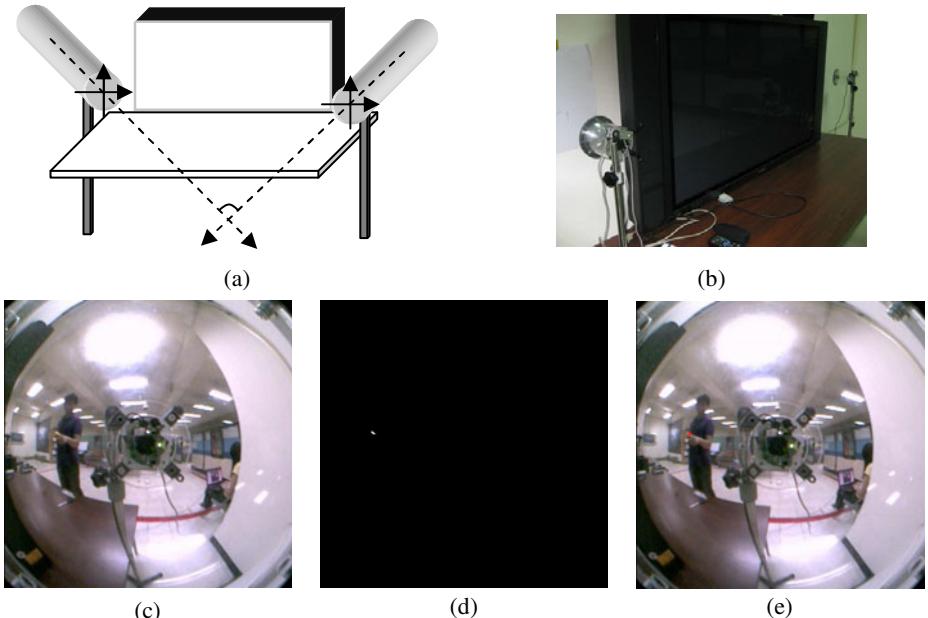


Fig. 1. Configuration of proposed system. (a) An illustration. (b) Real system used in this study. (c) An omni-image taken by right camera of a user wearing a yellow finger cot. (d) Detected finger cot region in (c). (e) Centroid of detected region of (d) marked as a red point.

More specifically, in the third stage a user is asked to measure the distance between the two camera stands first. The result, together with the angle between the two optical axes calculated in the second stage, is used to derive an *analytic* formula for calculating the 3D data of a feature point on the user's hand. Note that the two cameras connected in a line are assumed to be parallel to the console table surface. This requirement can be achieved by an in-field user by adjusting the two length-adjustable rods to an identical height. Additionally, it is assumed that each camera's optical axis is perpendicular to the vertical rod of the camera stand so that the axis is always parallel to the floor.

Before introducing the details of the three stages, we give a brief review of the structure of the omni-camera and the associated coordinate systems used in this study. As shown in Fig. 2, the type of omni-camera is catadioptric with the mirror being of a hyperboloidal shape, which may be described in the camera coordinate system (CCS) as:

$$\frac{R^2}{a^2} - \frac{(Z-c)^2}{b^2} = -1, \quad R = \sqrt{X^2 + Y^2},$$

where $a^2 + b^2 = c^2$. Specifically, the focal point O_m of the mirror is taken to be the origin $(0, 0, 0)$ of the CCS and the camera lens center O_c is located at $(0, 0, 2c)$ in the CCS.

As shown in Fig. 2(b), by triangulation, the camera focal length f may be computed as:

$$f = D \frac{r}{R}, \quad (1)$$

where D , R , and r respectively are the distance from the camera lens center to the mirror base center, the radius of the mirror base in the real-world space, and the radius of the mirror base in the image. Note that the mirror base is of a circular shape. Subsequently we will call the left and right cameras as Cameras 1 and 2, and their CCSs as CCS 1 and 2, respectively. A sketch of the proposed method is described in the following.

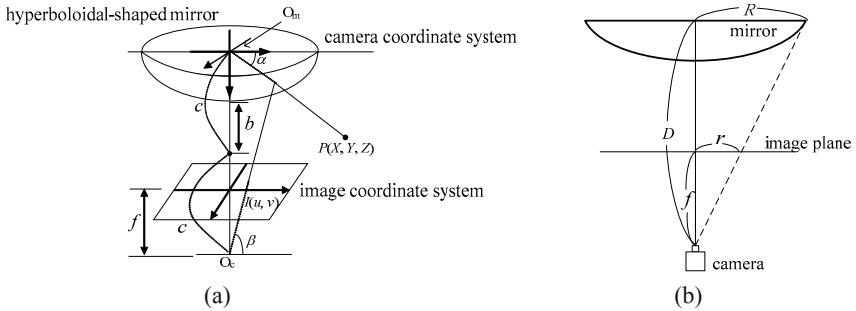


Fig. 2. Illustration of used omni-directional camera. (a) Camera and hyperboloidal-shaped mirror structure and coordinate systems. (b) Relationship between mirror and image plane.

Algorithm 1. Sketch of the proposed method.

Stage 1. Calibration of camera and mirror parameters.

- Step 1. Select a landmark point P in the real-world space.
- Step 2. Perform the following steps to calibrate two parameters of Camera 1.
 - 2.1 Measure manually the mirror base radius R_1 and the distance D_1 between the lens center and the mirror base center in the real-world space.
 - 2.2 Measure manually the location of landmark point P in CCS 1.
 - 2.3 Take an image I_1 of P with Camera 1 and extract the image coordinates (u_1, v_1) of the corresponding pixel p_1 from I_1 .
 - 2.4 Detect the boundary of the mirror base in I_1 and compute accordingly the mirror base radius r_1 .
 - 2.5 Use Eq. (1) to compute the focal length f_1 of Camera 1.
 - 2.6 With the location measured in Step 2.2, use a set of analytic formulas (derived later in Sec. 3) to compute the parameter $g_1 = c_1/b_1$, called *mirror distance ratio*, where b_1 and c_1 are the shape parameters of the mirror of Camera 1.
- Step 3. Take an image I_2 of P with Camera 2 and perform similarly steps to obtain the focal length f_2 and the mirror distance ratio g_2 of Camera 2.

Stage 2. Computing the optical axis angle by using a line feature on the console table.

- Step 4. Place the two camera stands at the left and right sides of the console table with the camera mirror facing forward to the user in front of the table.
- Step 5. Take an image of the line feature L of the front boundary of the console table by Camera 1 and extract its pixels.
- Step 6. Calculate two parameters (A_1, B_1) related to the direction of L from the viewpoint of Camera 1 by a Hough transform technique proposed in Sec. 4 with A_1 and B_1 to be derived in Sec. 4.
- Step 7. Take an image of the same line feature L by Camera 2, extract its pixels, and calculate two parameters (A_2, B_2) related to L from the viewpoint of Camera 2 by the same operations as Steps 5 and 6.
- Step 8. Calculate the angle θ between the optical axes of the two cameras in terms of A_1, B_1, A_2 , and B_2 by an equation derived in Sec. 4 (Eq. (12)).

Stage 3. Acquisition of 3D data of a feature points in the application activity.

- Step 9. Measure manually the distance d between the two cameras.
- Step 10. Take the images of a feature point P on the user hand by both cameras, and extract the pixels p_1 and p_2 corresponding to P in the images with coordinates (u_1, v_1) and (u_2, v_2) , respectively.
- Step 11. Calculate the 3D position of P with respect to CCS 1, using the image coordinates (u_1, v_1) and (u_2, v_2) as well as the angle θ calculated in Stage 2.

3 Proposed Techniques for Camera Parameter Calibration

The relation between the camera coordinates (X, Y, Z) of a space point P and the image coordinates (u, v) of its corresponding projection pixel p in an omni-image I as depicted in Fig. 2(a) may be described [9-12] by:

$$\tan \alpha = \frac{(b^2 + c^2) \sin \beta - 2bc}{(b^2 - c^2) \cos \beta}, \quad (2)$$

$$\cos \beta = \frac{r}{\sqrt{r^2 + f^2}}, \quad \sin \beta = \frac{f}{\sqrt{r^2 + f^2}}, \quad (3)$$

$$\tan \alpha = \frac{Z}{\sqrt{X^2 + Y^2}}, \quad (4)$$

where $r = \sqrt{u^2 + v^2}$. The angle α will be called the *elevation angle* of P . Also, according to the rotational invariance property of the omni-camera [12], we have the following equalities for describing the *azimuth angle* θ of p with respect to the u -axis in I and also of P with respect to the X -axis in the CCS:

$$\cos \theta = \frac{X}{\sqrt{X^2 + Y^2}} = \frac{u}{\sqrt{u^2 + v^2}}, \quad \sin \theta = \frac{Y}{\sqrt{X^2 + Y^2}} = \frac{v}{\sqrt{u^2 + v^2}}. \quad (5)$$

Traditionally, it is required to calibrate the three camera parameters b, c , and f for general applications. In this study, alternatively we adopt a more efficient way to compute first the focal length f by Eq. (1), which utilizes the easily-measurable mirror size and the distance between the camera and the mirror base, as described previously. And we then, by using the facts described by Eqs. (2) and (3) above, rewrite Eq. (4) to be

$$\tan \alpha = \frac{(g^2 + 1)\sin \beta - 2g}{(g^2 - 1)\cos \beta} \quad (6)$$

where g is defined as $g = c/b$. Eq. (6) may be solved to be

$$g = \frac{-1 \pm \sqrt{1 + (\tan \alpha)^2(\cos \beta)^2 - (\sin \beta)^2}}{\tan \alpha \cos \beta - \sin \beta}. \quad (7)$$

If we have a landmark point with known image coordinates (u, v) and known camera coordinates (X, Y, Z) , then we can calculate the parameter g by Eqs. (3), (4), and (7). In real situations, measurement errors and noise might occur, which result in inaccuracy of the computed value of g . To overcome this difficulty, one can use more than one landmark point to obtain a more accurate result by using non-linear regression or any other curve-fitting technique to get a better value of g fitting to Eq. (6).

4 Proposed Technique to Calculate the Angle between Optical Axes

In this section, we first propose a new technique in Sec. 4.1 to detect a space line in an omni-image and derive the two previously-mentioned parameters A and B to describe the line. The technique is then applied to the line feature of the front console table boundary. Finally, we propose a technique in Sec. 4.2 to calculate the angle between the optical axes of the two omni-cameras, which can be used for 3D feature point data computation.

4.1 Proposed Technique to Detect a Space Line in an Omni-Image

Given a space line L with a point $P_0 = (X_0, Y_0, Z_0)$ on the line and with the direction vector $v_L = (v_x, v_y, v_z)$, the equation of L may be written according to 3D geometry as

$$(X, Y, Z) = (X_0 + \lambda v_x, Y_0 + \lambda v_y, Z_0 + \lambda v_z)$$

where λ is a parameter. As shown in Fig. 3, let S be the space plane going through line L and the mirror base center $O_m = (0, 0, 0)$, and let $N_S = (l, m, n)$ be the normal vector of S . Then, any point P' on S with camera coordinates (X, Y, Z) and the mirror base center point O_m together form a vector $V_m = (X - 0, Y - 0, Z - 0) = (X, Y, Z)$ which is perpendicular to the normal vector $N_S = (l, m, n)$, so that the inner product of V_m and N_S becomes zero, leading the following equality:

$$N_S \cdot V_m = lX + mY + nZ = 0 \quad (8)$$

where “.” means the inner product of two vectors. Combining Eq. (8) with Eqs. (4) and (5) and performing some simplifications, we get

$$lR\cos\theta + mR\sin\theta + nR\tan\alpha = 0,$$

where $R = \sqrt{X^2 + Y^2}$. Dividing both sides of the above equality by R , we get

$$l\cos\theta + m\sin\theta + n\tan\alpha = 0,$$

which leads to

$$A \cos \theta + B \sin \theta = -\tan \alpha, \quad (9)$$

with $A = l/n$ and $B = m/n$. Note that the normal $N_S = (l, m, n)$ of plane S now may be expressed alternatively as $N_S = (A, B, 1)$.

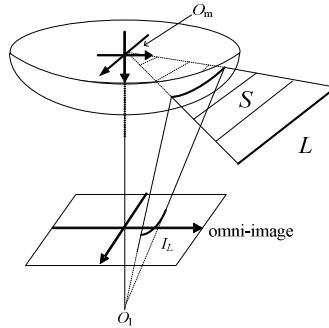


Fig. 3. Illustration of a space line projected on an omni-image

Eq. (9) describes the projection of a space line onto the omni-image in terms of the azimuth angle θ and the elevation angle α . The equation may be used, as found in this study, to detect the space line projection in an omni-image by a *simple* Hough transform as described next.

Algorithm 2. *Computing the parameters for detecting a space line in an omni-image.*

Input: The line feature l in an omni-image I (in the form of detected edge points) corresponding to a space line L .

Output: The values of the two parameters A and B in Eq. (9) which describes L .

Step 1. Extract the feature points of the line feature l in the omni-image I by an edge point detection algorithm.

Step 2. Set up a 2D Hough space with two parameters A and B and set all the initial cell values to be zero.

Step 3. Perform the following steps to each pixel of the line feature l with image coordinates (u, v) in I :

3.1 Compute $\cos \beta$, $\sin \beta$, $\cos \theta$, and $\sin \theta$ by Eq. (5) using u and v .

3.2 Compute $\tan \alpha$ by Eq. (6) using the values of $\cos \beta$ and $\sin \beta$, as well as g where the computation of g using landmark points was described at the end of Sec. 3.

3.3 For each cell at parameters (A, B) , if $\cos \theta$, $\sin \theta$, $\tan \alpha$, A , and B satisfy Eq. (9), then increment the cell value by one.

Step 4. Detect the peak cell value in the Hough space and take the corresponding parameters (A, B) as output.

4.2 Calculation of the Angle between the Two Optical Axes

Using Algorithm 2 above we can obtain the parameter pairs (A_1, B_1) and (A_2, B_2) to describe a space line L in CCS 1 and CCS 2, respectively. Let the directional vector of the space line L be $v_L = (v_x, v_y, v_z)$ again, and let the origin of CCS 1 be denoted as O_1 .

Also, let S_1 be a space plane going through line L and O_1 . As derived in Sec. 4.1, the normal vector of S_1 may be described by $n_1 = (A_1, B_1, 1)$. Since S_1 goes through line L , v_L and n_1 are perpendicular, meaning that

$$v_L \cdot n_1 = v_x A_1 + v_y B_1 + v_z = 0. \quad (10)$$

Furthermore, since the space line L is parallel to the xz -plane of CCS 1 and CCS 2 as assumed (see Sec. 2), we have

$$v_y = 0. \quad (11)$$

Combining Eqs. (10) and (11), we have

$$v_x A_1 + v_z = 0.$$

Taking v_x be 1 without loss of generality, we get the directional vector of L as $v_L = (v_x, v_y, v_z) = (1, 0, -A_1)$. Furthermore, as depicted in Fig. 2(a), the optical axis of either camera is parallel to the Z -axis of either CCS. Referring to Fig. 4, we see that the angle between the optical axis of Camera 1 and the space line L can be computed as

$$\theta_1 = \sin^{-1} \left(\frac{-A_1}{\sqrt{1+A_1^2}} \right).$$

Similarly, the angle between the optical axis of Camera 2 and the space line L can be computed as

$$\theta_2 = \sin^{-1} \left(\frac{-A_2}{\sqrt{1+A_2^2}} \right).$$

Accordingly, as depicted in Fig. 4, the angle θ between the two optical axes may be computed as

$$\theta = \theta_1 - \theta_2 = \sin^{-1} \left(\frac{-A_1}{\sqrt{1+A_1^2}} \right) - \sin^{-1} \left(\frac{-A_2}{\sqrt{1+A_2^2}} \right). \quad (12)$$

The value of θ derived above may be used to compute the 3D data of a feature point, as discussed in the following section.

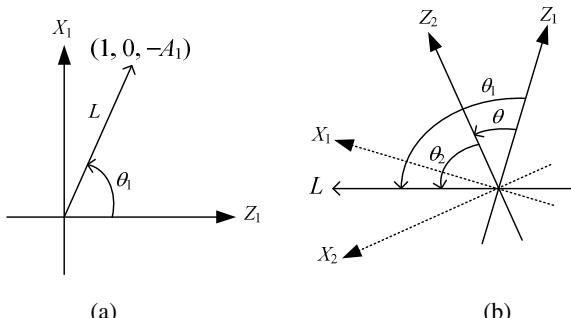


Fig. 4. Angles between space line L and optical axes Z_1 and Z_2 . (a) Directional vector of a space line L in CCS 1. (b) Angle between two optical axes.

5 Proposed Technique for Calculating 3D Data of Feature Points

To compute the 3D data of a space point $P = (X, Y, Z)$ from the two omni-images taken by the cameras as described in Stage 3 of Algorithm 1, let the projection of $P = (X, Y, Z)$ in an omni-image taken by Camera 1 be a pixel p_1 located at image coordinates (u_1, v_1) . From Eqs. (4) and (5), we have

$$X = R\cos\theta_1, Y = R\sin\theta_1, Z = R\tan\alpha_1 \quad (13)$$

where $\cos\theta_1$, $\sin\theta_1$, and $\tan\alpha_1$ can be computed by Eqs. (5) and (6). Eq. (13) can be viewed as one describing a line L_1 going through the origin $P_0 = (0, 0, 0)$ of CCS 1 with directional vector $v_1 = (\cos\theta_1, \sin\theta_1, \tan\alpha_1)$, and R being a parameter. Note that the coordinates (X, Y, Z) are with respect to CCS 1.

Similarly, let the feature point P be located at (X', Y', Z') in CCS 2, and let its projection in an omni-image taken by Camera 2 be a pixel p_2 located at image coordinates (u_2, v_2) . Then, similarly to the derivation of Eq. (13) we can obtain:

$$X' = R\cos\theta_2, Y' = R\sin\theta_2, Z' = R\tan\alpha_2. \quad (14)$$

With the aid of the angle θ between the two optical axes and the distance d between the two cameras (assumed to be known as mentioned before), we may transform the coordinates (X', Y', Z') in CCS 2 to the coordinates (X, Y, Z) in CCS 1 by the use of a rotation matrix M and a translation vector T described in the following:

$$T = (d, 0, 0), \quad M = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix},$$

or equivalently, we may transform (X', Y', Z') to (X, Y, Z) by the following equation:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = M \begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} + T = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} \begin{bmatrix} R\cos\theta_2 \\ R\sin\theta_2 \\ R\tan\alpha_2 \end{bmatrix} + \begin{bmatrix} d \\ 0 \\ 0 \end{bmatrix}$$

which can be viewed as a line L_2 going through a point $P_2 = (d, 0, 0)$ and with directional vector v_2 described as

$$v_2 = (\cos\theta\cos\theta_2 - \sin\theta\tan\alpha_2, \sin\theta_2, \sin\theta\cos\theta_2 + \cos\theta\tan\alpha_2).$$

We now have two space lines L_1 and L_2 going through point P . If everything including the system setup and the camera calibration is accurate without any error, these two lines should intersect precisely at one point P . But unavoidably, errors always exist, and so we estimate in this study the coordinates of point P as those of the mid-point of the shortest line segment between L_1 and L_2 .

6 Experimental Results

A series of experiments have been conducted with a calibration board to test the precision of the proposed system. The board is filled with grid points as shown in Fig. 5. We used nine of them on the board as landmark points in the experiments.

In an experiment, we put the two camera stands to face forward precisely, with the angle θ between the two camera optical axes being zero. Next, we measured manually the real 3D data of the nine landmark points. Then, we used the proposed method to compute the coordinates of each of the nine points in CCS 1. We define the *error ratio* of each point P_i at (X_i, Y_i, Z_i) , measured or computed, as

$$\frac{\sqrt{(real\ X_i - computed\ X_i)^2 + (real\ Y_i - computed\ Y_i)^2 + (real\ Z_i - computed\ Z_i)^2}}{d}$$

where d is the distance from P_i to the origin of CCS 1. The experimental results are summarized as in Table 1, from which we can see that the computed error ratios are all smaller than 2%, and the average of them is 1.39%.

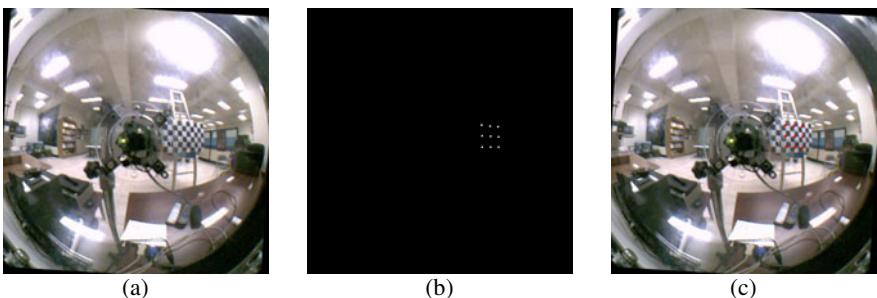


Fig. 5. Calibration board used in experiments. (a) Omni-image. (b) Nine calibration points on the board. (c) Calibration points superimposed on the image and shown in red.

Table 1. Error ratios of 3D data computation when two optical axes are mutually parallel

real X (cm)	computed X (cm)	real Y (cm)	computed Y (cm)	real Z (cm)	computed Z (cm)	error ratio
94.6	95.504	9	9.252	137	137.491	0.64%
94.6	96.658	29	28.553	137	138.101	1.41%
94.6	96.12	-11	-10.829	137	138.315	1.21%
74.6	76.413	9	9.333	137	138.667	1.59%
74.6	76.854	29	28.655	137	138.431	1.70%
74.6	77.114	-11	-9.792	137	137.776	1.85%
114.6	115.728	9	8.816	137	138.923	1.25%
114.6	115.795	29	28.874	137	139.712	1.64%
114.6	115.721	-11	-10.901	137	138.861	1.22%
average						1.39%

In another experiment, we simulated the situation that a user sets up the system beside the console table with unintended imprecision: the two camera stands do not face forward precisely. For this, we manually rotated the two camera stands with an *unknown* angle θ between the two optical axes within a range of about -5° to $+5^\circ$. Then, we used the proposed method to automatically detect the value of θ and used it to compute 3D feature point data. Specifically, each image pair taken by the two camera stands was processed according to steps in Stage 2 of Algorithm 1 to obtain the angle θ between the two optical axes. An example of experimental results is shown in

Fig. 6, where Fig. 6(a) is an image taken by Camera 1, Fig. 6(b) is the result of edge detection, Fig. 6(c) shows the edges in a region of interest (ROI), Fig. 6(d) is the parameter space of (A, B) produced by the Hough transform, Fig. 6(e) shows the projection of the detected line on the ROI, and Fig. 6(f) shows the projection of the detected line into Fig. 6(a). These results say that the Hough detection result is precise enough for real applications.

Table 2. Average error ratios obtained under imprecise system setup

Computed angle θ between optical axes	error ratio 1	error ratio 2	error ratio 3	error ratio 4	error ratio 5	error ratio 6	error ratio 7	error ratio 8	error ratio 9	average error ratio
-4.21	2.17%	1.83%	2.74%	1.77%	1.87%	1.88%	0.44%	0.89%	1.00%	1.62%
-4.10	0.60%	0.08%	0.53%	0.79%	0.91%	1.58%	0.50%	0.96%	1.07%	0.78%
-3.80	1.63%	1.26%	1.56%	1.26%	1.37%	1.56%	0.63%	0.96%	1.07%	1.25%
-3.16	0.76%	0.19%	1.34%	1.41%	1.44%	1.98%	0.92%	0.42%	0.54%	1.00%
-1.08	1.45%	1.72%	1.43%	2.77%	2.74%	2.73%	1.89%	2.26%	1.28%	2.03%
0.79	2.86%	2.56%	2.88%	2.18%	2.29%	3.31%	2.61%	2.34%	2.65%	2.63%
1.93	4.43%	3.60%	3.91%	3.18%	3.88%	4.33%	3.59%	3.32%	3.63%	3.76%
2.79	2.93%	2.70%	2.37%	2.13%	2.30%	2.68%	2.27%	2.58%	2.84%	2.53%
3.93	2.10%	1.90%	2.15%	1.20%	1.39%	1.88%	2.11%	1.88%	2.15%	1.86%
5.94	2.37%	2.19%	1.83%	1.40%	1.60%	2.05%	2.41%	2.19%	2.45%	2.05%
7.66	0.44%	0.49%	0.49%	1.48%	1.45%	1.70%	0.63%	1.04%	0.71%	0.94%
average										1.86%

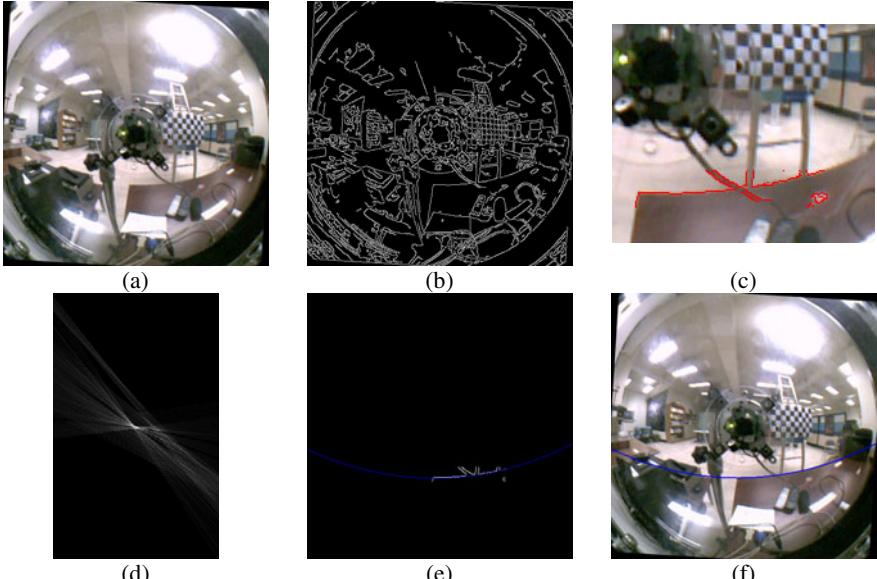


Fig. 6. An experimental result for Stage 2 of Algorithm 1. (a) Omni-image. (b) Edge detection result. (c) Edges in ROI. (d) Hough space. (e) Projection of peak cell with parameters $A = 0.035$ and $B = 1.395$ onto ROI. (f) Projection of detected line onto (a).

After deriving the angle between the optical axes, we measured the distance between two camera stands. Then, we use the nine points on the calibration board as used in Experiment 1 to test the precision of the proposed method in its capability of adaptation to imprecise system setups. The results are summarized in Table 2, from which we see that the average error ratios are still small and the overall average of them is 1.86%.

7 Conclusions

A new two-omni-camera system for 3D vision applications and a method for adapting it to imprecise camera setups have been proposed. First, a novel technique which uses a set of analytic formulas to calibrate omni-directional camera parameters has been proposed. Also, a technique for automatic detection of the angle between the two optical axes of the two cameras resulting from an imprecise camera setup has been proposed. Accordingly, a technique to adapt the system to the imprecise setup has been proposed as well, allowing the system to be able to conduct 3D feature data computation under the imprecise system configuration. Experimental results show that the proposed system and its adaptation capability yield 3D data computation results with average error ratios smaller than 3%, demonstrating the feasibility of the system for general applications.

References

1. Sogo, T., Ishiguro, H., Trivedi, M.M.: N-Ocular Stereo for Real-Time Human Tracking. In: Benosman, R., Kang, S.B. (eds.) *Panoramic Vision*, pp. 359–375. Springer, Berlin (2001)
2. Doubek, P., Svoboda, T.: Reliable 3D reconstruction from a few catadioptric images. In: *IEEE Workshop on Omni-directional Vision*, pp. 71–78 (2002)
3. Zhu, Z., Rajasekar, K., Riseman, E., Hanson, A.: Panoramic virtual stereo vision of cooperative mobile robots for localizing 3D moving objects. In: *IEEE Workshop on Omnidirectional Vision*, pp. 29–36 (2000)
4. Jeng, S.W., Tsai, W.H.: Analytic image unwarping by a systematic calibration method for omni-directional cameras with hyperbolic-shaped mirrors. *Image and Vision Computing* 26(5), 690–701 (2008)
5. Wu, C.J., Tsai, W.H.: Unwarping of images taken by misaligned omni-cameras without camera calibration by curved quadrilateral morphing using quadratic pattern classifiers. *Optical Engineering* 48(8), Article no. 087003 1–11 (2009)
6. Fabrizio, J., Tarel, J.-P., Benosman, R.: Calibration of panoramic catadioptric sensors made easier. In: *IEEE Workshop on Omni-directional Vision*, pp. 45–52 (2002)
7. Mei, C., Rives, P.: Single View Point Omnidirectional Camera Calibration from Planar Grids. In: *IEEE International Conference on Robotics and Automation*, pp. 3945–3950 (2007)
8. Wu, C.J., Tsai, W.H.: An omni-vision based localization method for automatic helicopter landing assistance on standard helipads. In: *International Conference on Computer and Automation Engineering*, Singapore, pp. 327–332 (2010)
9. Gaspar, J., Winters, N., Santos-Victor, J.: Vision-based navigation and environmental representations with an omni-camera. *IEEE Transactions on Robotics and Automation* 16(6), 890–898 (2000)

10. Mashita, T., Iwai, Y., Yachida, M.: Calibration method for misaligned catadioptric camera. *IEICE Transactions on Information & Systems* E89-D(7), 1984–1993 (2006)
11. Ukida, H., Yamato, N., Tanimoto, Y., Sano, T., Yamamoto, H.: Omni-directional 3D Measurement by Hyperbolic Mirror Cameras and Pattern Projection. In: *IEEE Conference on Instrumentation and Measurement Technology*, pp. 365–370 (2008)
12. Jeng, S.W., Tsai, W.H.: Using pano-mapping tables to unwarping of omni-images into panoramic and perspective-view Images. *IET Image Processing* 1(2), 149–155 (2007)

3D Face Recognition Based on Local Shape Patterns and Sparse Representation Classifier

Di Huang¹, Karima Ouje¹, Mohsen Ardabilian¹, Yunhong Wang², and Liming Chen¹

¹ LIRIS Laboratory, CNRS 5205, Ecole Centrale Lyon, 69134, Lyon, France

² School of Computer Science and Engineering, Beihang University, 100091, Beijing, China
{di.huang,karima.ouje,mohsen.ardabilian,liming.chen}@ec-lyon.fr,
yhwang@buaa.edu.cn

Abstract. In recent years, 3D face recognition has been considered as a major solution to deal with these unsolved issues of reliable 2D face recognition, i.e. illumination and pose variations. This paper focuses on two critical aspects of 3D face recognition: facial feature description and classifier design. To address the former one, a novel local descriptor, namely Local Shape Patterns (LSP), is proposed. Since LSP operator extracts both differential structure and orientation information, it can describe local shape attributes comprehensively. For the latter one, Sparse Representation Classifier (SRC) is applied to classify these 3D shape-based facial features. Recently, SRC has been attracting more and more attention of researchers for its powerful ability on 2D image-based face recognition. This paper continues to investigate its competency in shape-based face recognition. The proposed approach is evaluated on the IV² 3D face database containing rich facial expression variations, and promising experimental results are achieved which prove its effectiveness for 3D face recognition and insensitivity to expression changes.

Keywords: 3D face recognition, local descriptor, Local Shape Patterns (LSP), Sparse Representation Classifier (SRC).

1 Introduction

Human face is potentially the best biometrics for verification and identification tasks, since it is non instructive, contactless and socially well accepted [1]. The past several years have witnessed the tremendous research efforts first focused on 2D face images [2] and recently on 3D face models [3]. In despite of great progress achieved so far in the field [2], 2D face image, as one of biometrics, does not remain reliable enough [4], especially in the presence of pose and lighting variations [5]. Along with progress in 3D imaging system, 2.5D or 3D face scans have emerged as a major solution to handle these unsolved issues of 2D face recognition [3, 6].

Zhao et al. [2] categorized 2D image-based face recognition techniques into three main approaches: the holistic; feature-based and hybrid ones. This taxonomy can also be extended to 3D model-based face recognition. E.g., the holistic category includes ICP (Iterative Closest Point) based matching [11], annotated deformable model [12], isometry-invariant description [13] etc. The matching scheme based on holistic

features has tolerance to noise, but it always requires accurate normalization with respect to poses and scales. Moreover, it proved sensitive to facial expression variations and partial occlusions. Feature-based schemes compare local descriptive points or regions of 3D face scans and have been explored by several tasks in the literature, containing the point signature approach [14] and multi-modal local feature-based matching [15]. Feature-based matching has the potential advantages of being robust to facial expression, pose variations and even to partial occlusions, while its downside is the difficulty extracting sufficient repeatable informative local features from similar and smooth 3D facial surfaces. There also exist some papers presenting hybrid matching which combines global features with local ones: Region-ICP matching [16], multiple region-based matching [17], component and morphable model-based matching [18]. However, it risks inheriting both types of shortcomings: sensitivity to pose changes, difficulty generating sufficient stable descriptive features, etc.

Generally, a robust face recognition system involves two crucial aspects: facial feature description and classifier design [7]. Facial description is to derive a set of features from original images for representing faces. If features are inadequate, even the best classifier will fail to achieve an accurate recognition result. Therefore, it is critical to extract discriminative features for facial representation. “Good” facial features are desired to have following properties [7]: first, can tolerate the within-class variations while discriminate different classes well; second, can be easily extracted from raw images to allow fast processing; finally, lie in a space with a moderate dimensionality to avoid high computation cost. On the other hand, classifier design is to seek proper methods or schemes to compare facial features, and the classifier is expected to possess strong discriminative power on these features.

In 3D domain, many typical features have been applied for facial description. For instance, Samir et al. [27] proposed to represent a facial surface by using the union of the level curves of a depth function. Chua et al. [14] introduced point signature for 3D face representation. Tanaka et al. [28] adopted an Extended Gaussian Image (EGI) as facial features. Xu et al. [29] extracted Gabor features from range faces. Lu et al. [11] computed Shape Index (SI) to describe local characters of facial shape. Li et al. [20] attempted to use Local Binary Patterns (LBP) for 3D face recognition, and later, their performance was improved by Huang et al. [21] who extended LBP to 3DLBP. When regarding the classification step, Nearest Neighbor (NN), Support Vector Machine (SVM), and Chi-Square distance are very popular.

According to the two critical aspects above, in this paper, we proposed a novel local 3D shape descriptor, namely Local Shape Patterns (LSP), to represent local shape attributes. LSP operator reproduces the way as computing LBP features, it thus works very efficiently. Besides the differential structure provided by LBP operator, LSP also extract orientation information, which describes 3D facial surfaces more comprehensively when compared with LBP and even Shape Index [8]. Meanwhile, more recently, Sparse Representation Classifier (SRC) has attracted more and more attention of researchers for its powerful capacity on 2D image based face recognition. This paper further explores its competency in the classification of 3D shape based facial features. The proposed approach is evaluated on the IV² 3D face database which contains rich facial expression changes, and experimental results illustrate that it is effective for 3D face recognition and insensitive to expression variations.

The remainder of this paper is organized as follows. The proposed approach overview is depicted in section 2. R-ICP based 3D facial surface registration is described in section 3. The Local Shape Pattern (LSP) operator is presented in section 4; while section 5 introduces SRC. The experimental results are shown and analyzed in section 6, and section 7 concludes the paper.

2 The Approach Overview

This paper presents a feature-based approach for 3D face recognition, which consists of three main steps: 3D registration, local shape description, as well as classification. Its framework is shown in Fig.1. For face registration, Region-based Iterative Closest Point (R-ICP) [16] is adopted that only uses the rigid facial region, and hence it can achieve more accurate pose correction than ICP does. After all the 3D face models in both the gallery and probe sets are registered to correct pose variations, their corresponding range face images are generated. Then the LSP operator works on each range image to extract local shape features containing both differential structure and orientation information, which describes 3D facial surface comprehensively. Finally, SRC is applied to classify the LSP-based facial features and give the final decision.

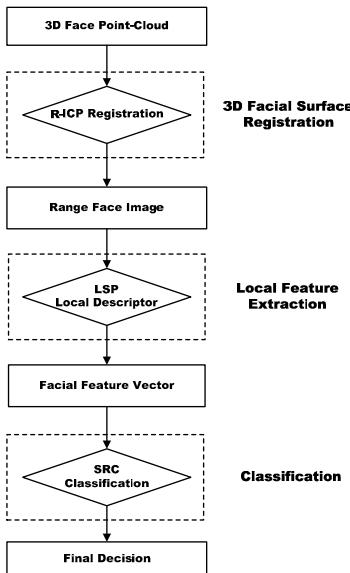


Fig. 1. The framework of the proposed approach

3 3D Face Registration Using R-ICP

3D face registration is an important step to correct pose variations in 3D face analysis, and Iterative Closest Point (ICP) algorithm [10] is considered as a popular and effective alternative.

The ICP methodology is an iterative procedure minimizing the Mean Square Error (MSE) between points in the reference model and the closest vertices, respectively, in the test model. At each iteration of this algorithm, geometric transformation which best aligns the two 3D face models is computed. This procedure operates until convergence, and always converges monotonically to a local minimum. Convergence to a global minimum needs a good initialization.

For this reason, a coarse-to-fine strategy is adopted which consists of a coarse step, which approximates the rigid transformation between models and bring them closer; and a further fine step calculating the minimal distance which converges to a minima starting from the initial solution (see Fig. 2).

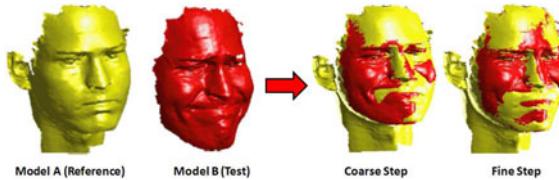


Fig. 2. The ICP-based 3D facial surface registration process

ICP extracts and computes rotation and translation which characterize the rigid deformation between the reference and test model. However, facial expressions cause the non-rigid deformations that are unable to be modeled just by rotation and translation. It is thus necessary to identify and distinguish the dynamic region which is sensitive to expression variations from a static one.



Fig. 3. The face region segmentation

In this paper, an improved version of ICP, namely Region-based Iterative Closest Point (R-ICP) [16], is introduced for registration which only makes use of the static region. The static and dynamic regions are fixed by the segmentation method already presented in [19], and Fig. 3 shows an example.

After 3D registration, the frontal range face images are generated for the following feature extraction step (see Fig. 4).



Fig. 4. Range face examples

4 Local Feature Extraction

In recent years, local feature based approaches have attracted much more attention than the holistic ones for facial image analysis. Among the off-the-shelf local descriptors, Local Binary Patterns (LBP) [19] and Scale Invariant Feature Transform (SIFT) [20] are especially popular and powerful. The former one extracts local texture micro-patterns by comparing each pixel with its neighboring pixels, holding the properties of tolerance regarding monotonic illumination variations and its computational simplicity; while the latter one provides excellent performance in the context of matching and recognition due to its invariance to scaling and rotations.

LBP and its variations have been extensively adopted for 2D face analysis during the last several years, since it was first used for face recognition [22]. In the domain of 3D face recognition, feature extraction often works on range face images. Since the range face image only displays geometric information, it is quite different from that of texture formed by light reflection. However, according to the definition, LBP can also describe local shape conditions, such as flat, concave and convex instead of bright or dark spot, edge in texture images. In the literature, Li et al. [20] used LBP directly on range images to achieve 3D face recognition. Later, in [21], Huang et al. proposed an extended LBP version, named 3DLBP, for the same target. Besides the information provided by LBP, 3DLBP also considers the exact value difference between the given pixel and its neighboring pixel, and it proved superior to LBP.

On the other hand, SIFT is not widely utilized for face recognition, and only a few papers applied SIFT to 2D face recognition. Since all the face images are cropped and normalized to the same size in preprocessing, the property of scale invariance in SIFT operator is not so critical for face recognition. Instead, in our opinion, its orientation information is still helpful to describe faces especially to represent 3D faces.

Therefore, inspired by LBP and SIFT methodology, a novel local descriptor of 3D facial surface, named Local Shape Patterns (LSP), is proposed to describe local shape attributes, which combines the advantages of LBP and SIFT and thus provides both the differential structure and orientation information. Moreover, it works very fast and is easy to be extended.

4.1 Local Binary Patterns

Specifically, the original LBP labels each pixel of a given image by thresholding in a 3×3 neighborhood. If the values of the neighboring pixels are not lower than that of the central pixel, their corresponding binary bits are assigned to 1; or to 0. A binary number is formed by concatenating all the eight binary bits, and the resulting decimal value is used for labeling. Fig. 5 gives an example of this process.

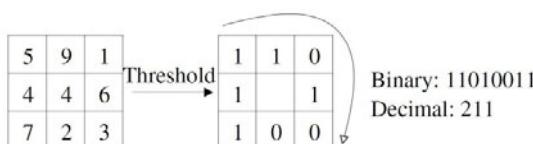


Fig. 5. An example of the original LBP operator

Formally, given a pixel at (x_c, y_c) , the derived LBP decimal value is:

$$LBP(x_c, y_c) = \sum_{n=0}^8 s(i_n - i_c) 2^n; \quad (1)$$

where n covers the eight neighbors of the central pixel, i_c and i_n are gray level values of the central pixel and its surrounding pixels respectively. Function $s(x)$ is compute as follows:

$$s(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (2)$$

The original LBP operator was extended later with various local neighborhood sizes to deal with different scales. The local neighborhood of LBP is defined as a set of sampling points evenly spaced on a circle which is centered at the pixel to be labeled. The sampling points that do not fall exactly on the pixels are expressed using bilinear interpolation, and thus allowing any value of radius and any number of points in the neighborhood. Fig.6 shows different LBP neighborhoods. The notation (P, R) denotes the neighborhood of P sampling points on a circle of radius R .

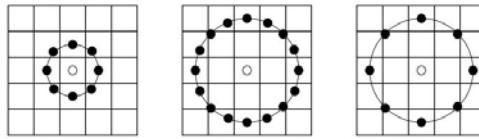


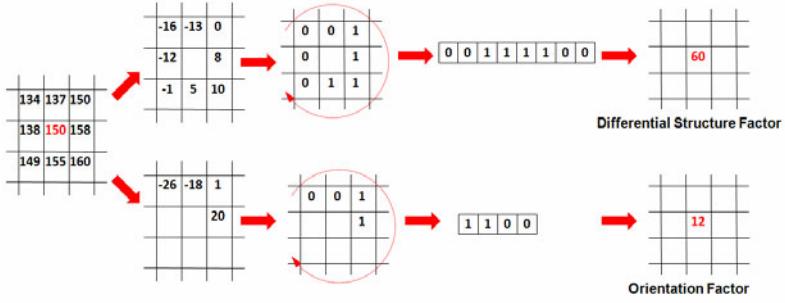
Fig. 6. Operator examples: circular (8, 1), (16, 2), and (8, 2)

According to Equa. (1), LBP is invariant to monotonic gray-scale transformations preserving pixel order in local neighborhoods; therefore, it has been considered one of the most effective and popular 2D texture descriptors. Based on definition (1), when it works on depth information of range images, the generated local binary patterns can also describe local geometry structures.

4.2 Local Shape Patterns

Similar with original LBP, the basic LSP operator also works on a 3x3 neighborhood. It not only extracts the differential structure as LBP does, but also provides the orientation as SIFT-based features. Differential structure is directly computed by LBP; the decimal value formed by the corresponding LBP code of the given pixel is the differential structure factor. With a 3x3 neighborhood, there are 256 (2^8) differential patterns. To calculate the orientation information, the first four neighboring pixels (also start from top-left in a clockwise direction, see Fig. 7) subtract the pixels in the same diameters of the circle centered at the given pixel, and then the decimal value of the 4-bit binary codes is orientation factor of the given pixel. There are thus 16 (2^4) orientation patterns. Fig. 7 shows an LSP illustration.

From Fig. 7, we can see that if two local points has the same differential structure, additional orientation information is helpful to distinguish them. LSP, thus, is more discriminative to describe local shape attributes than LBP. To handle shape attributes at different scales, LSP operator can be generalized to use neighborhoods of various sizes inheriting from that of LBP.

**Fig. 7.** The Illustration of LSP

Given a pixel at (x_c, y_c) , the resulting LSP can be expressed as:

$$LSP_{P,R}(x_c, y_c) = \{D_{P,R}(x_c, y_c), O_{P,R}(x_c, y_c)\} \quad (3)$$

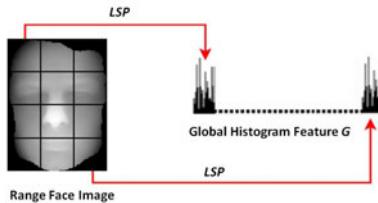
where D and O denotes the differential structure and orientation respectively. D and O are computed as follows:

$$D_{P,R}(x_c, y_c) = \sum_{n=0}^{P-1} s(i_n - i_c) 2^n \quad (4)$$

$$O_{P,R}(x_c, y_c) = \sum_{n=0}^{\left(\frac{P}{2}\right)-1} s\left(i_n - i_{n+\left(\frac{P}{2}\right)}\right) 2^n \quad (5)$$

where i_c and i_n are intensity values of the central pixel and P surrounding pixels in the neighborhood respectively, and function $s(x)$ uses the same definition in (2).

In order to describe a local shape region, histogram statistic is introduced as facial feature vector. In LSP operator, the histograms of differential structure and orientation can be combined in different ways, such as fusion of their individual histograms at the feature level or at the score level. Motivated by [23], we propose the strategy that for each orientation pattern o_i , a 1-D histogram h_i is generated by its corresponding differential structure patterns; all h_i are then concatenated to generate final histogram H .

**Fig. 8.** LSP based facial representation

4.3 LSP Based Facial Representation

The LSP based facial representation is achieved in a hybrid way that the entire range image is first divided into some non-overlapped regions, from which their local shape

features H are extracted; then concatenated based on facial component configuration to form a global feature G to represent the face (see Fig. 8).

5 SRC Classification

Sparse Representation for Signal Classification (SRSC) was first proposed to incorporate reconstruction properties, discriminative power and sparsity for robust classification [24]. A general classification method SRC for 2D face recognition was presented using a sparse representation computed by L1-minimization [25]. It always achieves high accuracy as lighting variation and occlusion occurs. Considering its high discriminative power, in this paper, we explore SRC to classify 3D features extracted from range face images.

k classes and n_i feature vectors, $v_{i,j} \in R_m$ are for training from i^{th} class, $i = \{1, 2, \dots, k\}$ and j is the index of the training sample, $j = \{1, 2, \dots, n_i\}$. All the training data from the i^{th} class are placed in a matrix $A_i = [v_{i,1}, v_{i,2}, \dots, v_{i,n_i}] \in R_{m \times n_i}$. A dictionary matrix A for k classes is developed by concatenating A_i , $i = 1, \dots, k$. A test pattern y can be represented as a linear combination of all n training samples ($n = n_i \times k$):

$$y = Ax \quad (6)$$

where x is an unknown coefficient vector. In (6), it is straightforward to note that only the entries of x that are non-zero correspond to the class of y . Equation (6) can be solved according to compressed sensing as long as the solution to (6) is known to be sufficiently sparse. An equivalent L1-norm minimization:

$$(L1) : x_l = \arg \min \|x\|_1 ; Ax = y \quad (7)$$

can be solved as a good approximation to Equa. (6). With the solution x_l of (7), we can compute the residual between a given probe face and each individual as:

$$R_i = \left\| y - \sum_{j=1}^k x_{l_{i,j}} v_{i,j} \right\|_2 \quad (8)$$

The identity of any given probe face is then determined as the one with the smallest residual R .

6 Experiments and Results

The proposed method is tested on the IV² 3D face dataset [26] possessing more facial expression than the well-known FRGC v2.0 does. IV² contains 50 subjects, each of which has six 3D face models. For each subject, the neutral expression model is employed as gallery; while the probe set consists of the other five face models with five facial expression variations, i.e. neutral, closed eyes, disgust, happy and surprise (see Fig. 4). In the probe set, there are 250 3D face models in total. To evaluate the degradation of performance when expression variations occur, the probe set is divided into two subsets: Subset I with 50 neutral ones while Subset II with 200 non-neutral ones. A preprocessing technique is adopted to remove spikes with the median filter and fill holes using cubic interpolation. All range faces are normalized to 200×150 pixels.

In our experiments, the LSP operator uses the parameter setting of 8 neighboring points and 2 pixel radius for local feature extraction. To prove the effectiveness of LSP based facial features, the results based on the original range image (R), differential structure (D) or orientation (O) only are provided. We also implement LBP [15], 3DLBP [16] as well as Shape Index (SI) for comparison. It should be noted that in [15], LBP was directly applied to extract features of 3D faces, and it actually achieves the same performance as that only based on differential structure. Moreover, since [16] used the same dataset, the accuracy is directly cited; while to illustrate the discriminative power of SRC on histogram based features of 3D faces, the corresponding results using Chi-Square distance, a well known classifier for histogram based features, are also shown. The results of feature level fusion (FF) and score level fusion (SF) aim to highlight the efficiency of the proposed combination of D and O.

In each range face, all the sub-regions are rectangle of size 25×25 pixels; and thus each image has 48 divisions. Sum rule is used for score level fusion. Rank-one recognition rates of different approaches are listed in Table.1.

Table 1. Rank-one face recognition rates of different methods on the IV2 3D face database

Approaches	Rank-one Face Recognition Rate
(1) D + Chi-Square [15]	0.756
(2) O + Chi-Square	0.576
(3) FF + Chi-Square	0.776
(4) SF + Chi-Square	0.764
(5) 3DLBP + Chi-Square [16]	0.796
(6) LSP + Chi-Square	0.804
(7) R + SRC	0.808
(8) SI + SRC [8]	0.816
(9) D + SRC [15]	0.880
(10) O + SRC	0.804
(11) FF + SRC	0.896
(12) SF + SRC	0.896
(13) 3DLBP + SRC [16]	0.900 (I: 0.900; II: 0.900)
(14) LSP + SRC	0.920 (I: 0.960; II: 0.910)
(15) Geodesic [12]	0.872 (I: 0.960; II: 0.850)

I: Neutral Probes; II: Non-Neutral Probes.

From Table.1, the results of LSP based 3D facial feature are better than that based on only D or O using both the classifiers, i.e. Chi-Square distance and SRC, showing that both differential structure and orientation are useful to describe the local shape attribute of 3D facial surfaces. With the two types of classifiers, LSP also achieves better performance than those of two fusion scheme: FF and SF, and it proves that the proposed combination method is effective. SRC outperforms Chi-Square distance for classifying all the histogram based 3D facial features, which illustrates the discriminative power of SRC on 3D depth information based features. Comparing the results of (7), (8), (9), (13) and (14), we can see LSP is superior to the others when describing local shape attributes. To sum up, both LSP and SRC are necessary to achieve the good final performance.

7 Conclusions

In recent years, 3D face recognition has been considered as a major solution to handle the unsolved issues of reliable 2D face recognition, i.e. illumination and pose changes. This paper discusses two important aspects in 3D face recognition, i.e., facial feature description and classifier design. To address the former one, a novel local descriptor, namely Local Shape Patterns (LSP), to represent local shape attributes more comprehensively, since LSP provides both the differential structure and orientation information. For the latter one, Sparse Representation Classifier (SRC) is introduced to classify 3D depth information based facial features.

The proposed method is evaluated on the IV² 3D face dataset, and the experimental performance not only illustrates the effectiveness of LSP for local shape description and the discriminative power of SRC for classifying these LSP-based 3D facial features, but also proves that our approach consisting of R-ICP 3D face registration, LSP based 3D facial feature extraction and SRC based classification is efficient for 3D face recognition and insensitive to facial expression variations.

In future work, we will evaluate the proposed approach with multi-scale scheme.

References

1. Jain, A.K., Ross, A., Prabhakar, S.: An Introduction to Biometric Recognition: Special Issue on Image- and Video-based Biometrics. *IEEE Transactions on Circuits and Systems for Video Technology* 14(1), 4–20 (2004)
2. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face Recognition: A Literature Survey. *ACM Computing Survey* 35(4), 399–458 (2003)
3. Bowyer, K.W., Chang, K., Flynn, P.J.: A Survey of Approaches and Challenges in 3D and Multi-modal 3D+2D Face Recognition. *Computer Vision and Image Understanding* 101(1), 1–15 (2006)
4. Abate, A.F., Nappi, M., Riccio, D., Sabatino, G.: 2D and 3D Face Recognition: A Survey. *Pattern Recognition Letters* 28(14), 1885–1906 (2007)
5. Phillips, P.J., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET Evaluation Methodology for Face Recognition Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(10), 1090–1104 (2000)
6. Scheenstra, A., Ruifrok, A., Veltkamp, R.C.: A Survey of 3D Face Recognition Methods. In: International Conference on Audio- and Video- based Biometric Person Authentication (2005)
7. Hadid, A., Pietikäinen, M., Ahonen, T.: A Discriminative Feature Space for Detecting and Recognizing Faces. In: IEEE International Conference on Computer Vision and Pattern Recognition (2004)
8. Koenderink, J.J., Doorn, A.J.: Surface Shape and Curvature Scales. *Image and Vision Computing* 10(8), 557–565 (1992)
9. Ojala, T., Pietikäinen, M., Maenpää, T.: Multiresolution Gray-scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence* 24(7), 971–987 (2002)
10. Besl, P.J., McKay, N.D.: A Method for Registration of 3-D Shapes. *IEEE Trans. Pattern Analysis and Machine Intelligence* 14(2), 239–256 (1992)

11. Lu, X., Jain, A.K., Colbry, D.: Matching 2. 5D Face Scans to 3D Models. *IEEE Trans. Pattern Analysis and Machine Intelligence* 28(1), 31–43 (2006)
12. Kakadiaris, I.A., Passalis, G., Toderici, G., Murtuza, M.N., Lu, Y., Karampatziakis, N., Theoharis, T.: Three-dimensional Face Recognition in the Presence of Facial Expressions: An Annotated Deformable Model Approach. *IEEE Trans. Pattern Analysis and Machine Intelligence* 29(4), 640–649 (2007)
13. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Three-dimensional Face Recognition. *International Journal of Computer Vision* 64(1), 5–30 (2005)
14. Chua, C., Han, F., Ho, Y.K.: 3D Human Face Recognition using Point Signature. In: *IEEE International Conference on Automatic Face and Gesture Recognition* (2000)
15. Mian, A.S., Bennamoun, M., Owens, R.: Keypoint Detection and Local Feature Matching for Textured 3D Face Recognition. *International Journal of Computer Vision* 79(1), 1–12 (2008)
16. Ouji, K., Amor, B.B., Ardabilian, M., Chen, L., Ghorbel, F.: 3D Face Recognition using Region-ICP and Geodesic Coupled Approach. In: *Conference on International Multimedia Modeling* (2009)
17. Mian, A.S., Bennamoun, M., Owens, R.: Region-based Matching for Robust 3D Face Recognition. In: *Conference on British Machine Vision* (2005)
18. Huang, J., Heisele, B., Blanz, V.: Component-based Face Recognition with 3D Morphable Models. In: *International Conference on Audio- and Video- Based Biometric Person Authentication* (2003)
19. Amor, B.B., Ardabilian, M., Chen, L.: Enhancing 3D Face Recognition by Mimic Segmentation. In: *International Conference on Intelligent Systems Design and Applications* (2006)
20. Li, S.Z., Zhao, C., Ao, M., Lei, Z.: Learning to Fuse 3D+2D based Face Recognition at Both Feature and Decision Levels. In: *International Workshop on Analysis and Modeling of Faces and Gesture* (2005)
21. Huang, Y., Wang, Y., Tan, T.: Combining Statistics of Geometrical and Correlative Features for 3D Face Recognition. In: *Conference on British Machine Vision* (2006)
22. Ahonen, T., Hadid, A., Pietikäinen, M.: Face Recognition with Local Binary Patterns. In: Pajdla, T., Matas, J.(G.) (eds.) *ECCV 2004. LNCS*, vol. 3021, pp. 469–481. Springer, Heidelberg (2004)
23. Chen, J., Shan, S., Zhao, G., Chen, X., Gao, W., Pietikäinen, M.: A Robust Descriptor based on Weber's Law. In: *IEEE International Conference on Computer Vision and Pattern Recognition* (2008)
24. Huang, K., Aviyente, S.: Sparse Representation for Signal Classification. In: *Annual Conference on Neural Information Processing Systems* (2006)
25. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust Face Recognition via Sparse Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(2), 210–227 (2009)
26. Colineau, J., D'Hose, J., Amor, B.B., Ardabilian, M., Chen, L., Dorizzi, B.: 3D Face Recognition Evaluation on Expressive Faces using the IV2 Database. In: *International Conference on Advanced Concepts for Intelligent Vision Systems* (2008)
27. Samir, C., Srivastava, A., Daoudi, M.: Three-dimensional Face Recognition using Shapes of Facial Curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(11), 1858–1863 (2006)
28. Tanaka, H.T., Ikeda, M., Chiaki, H.: Curvature-based Face Surface Recognition using Spherical Correlation — Principal Directions for Curved Object Recognition. In: *IEEE International Conference on Automatic Face and Gesture Recognition* (1998)
29. Xu, C., Li, S.Z., Tan, T., Quan, L.: Automatic 3D Face Recognition from Depth and Intensity Gabor Features. *Pattern Recognition* 42(9), 1895–1905 (2009)

An Effective Approach to Pose Invariant 3D Face Recognition

Dayong Wang, Steven C.H. Hoi, and Ying He

School of Computer Engineering, Nanyang Technological University, Singapore
`{S090023, chhoi, yhe}@ntu.edu.sg`

Abstract. One critical challenge encountered by existing face recognition techniques lies in the difficulties of handling varying poses. In this paper, we propose a novel pose invariant 3D face recognition scheme to improve regular face recognition from two aspects. Firstly, we propose an effective geometry based alignment approach, which transforms a 3D face mesh model to a well-aligned 2D image. Secondly, we propose to represent the facial images by a Locality Preserving Sparse Coding (LPSC) algorithm, which is more effective than the regular sparse coding algorithm for face representation. We conducted a set of extensive experiments on both 2D and 3D face recognition, in which the encouraging results showed that the proposed scheme is more effective than the regular face recognition solutions.

1 Introduction

Face recognition, an important biometrics technique, plays a critical role in many real-world multimedia applications. Despite being studied extensively in literature [1,9], existing face recognition techniques still suffer from a lot of challenges when being applied in real-world applications. In particular, many 2D face recognition approaches work excellently under well-controlled conditions (well-posed and good lighting), but their recognition accuracy often decreases considerably when handling real-world face recognition tasks where variations are common for pose, illumination and expression [13,18].

On the other hand, along with the advances of various 3D capture devices, 3D face recognition techniques are receiving more and more research attention [1,5]. The highly detailed 3D mesh data can capture rich information, which potentially provides much more clues to tackle some unsolved challenges in 2D face recognition tasks, especially for pose and illumination variations.

Following this direction, in this paper, we investigate a novel 3D face recognition scheme that addresses the open challenge of *Pose Invariant Face Recognition*(PIFR). In particular, we propose an effective approach to tackling the pose invariant 3D face recognition task, which is equipped with a set of effective 3D parametrization, alignment, and spares feature representation techniques.

Specifically, the main contributions of this paper include:

- We propose a new 3D face recognition approach to pose invariant face recognition, which employs a state-of-the-art 3D parameterization technique to resolve the challenge of pose invariant face alignment.
- We propose a new Locality Preserving Sparse Coding (LPSC) algorithm for facial image feature representation, which is empirically more effective for face recognition than the regular sparse coding method [10].

The rest of this paper is organized as follows. Section 2 reviews related work on PIFR and sparse coding. Section 3 presents an overview of the proposed PIFR system. Section 4 discusses a geometry-based face alignment approach by applying an effective 3D mesh parameterization technique. Section 5 presents the proposed novel LPSC algorithm. Section 6 presents an extensive set of empirical studies for performance evaluation, and Section 7 concludes this work.

2 Related Work

Below we review two major groups of related work: pose invariant face recognition and sparse coding techniques.

Pose Invariant Face Recognition. To attack a pose-invariant face recognition task, one possible remedy approach is to capture multi-view face images from each individual and estimate all the other possible pose positions. However, it is often not practical to collect multi-view images for each individual in real applications. As a result, the virtual view synthesis scenarios, which base on *2D pose transformation* or *3D face reconstruction*, are proposed to substitute the demand of real views from limited known views(i.e. only the frontal view in our framework) [17].

The 2D pose transformation schemes, such as active shape model (ASM) and active appearance models(AAM) [6], have been demonstrated to handle the PIFR problems effectively within small-scale pose variation. Unfortunately, they often fail for large-scale in-depth face rotation (i.e. larger than 45°) because of the image discontinuities. By utilizing the local feature instead of the whole image, some transformation algorithms could partially overcome the former limitation and further boost the performance. For example, Prince et al. [12] proposed a statistical linear model, called “tied” factor analysis model(TFA), which constructs a one-to-many mapping from the “identify” space to the observed image space with the pose-contingent linear transformation. Comparing with the state-of-the-art 3D face reconstruction approach, they achieved comparative experimental results with 14 manually-identified keypoints on each face. Their method however falls short in very intensive computational costs using local features, and their performance often highly depends on the benchmark point detection or even lots of manual labeling efforts.

The 3D face reconstruction schemes have the potentiality to overcome the pose variance challenge and achieve satisfactory results. However, the 3D reconstruction schemes are complex to implement and extremely computationally expensive because of the slow 3D face modeling process.

With the rapid improvement in 3D capture devices, face recognition techniques, based on 3D data directly, are receiving more and more research attention. It is potentially promising technique to overcome this challenge by exploiting the internally invariance of viewpoint and illumination in the 3D scanned data. Among these works, the techniques, such as extended Gaussian images(EGI), ICP matching, hausdorff distance matching and so on, are proposed for 3D shape recognition. Multi-model approaches, which combine the 2D and 3D results, are also developed to enhance the performance [5].

However, existing 3D face recognition methods rarely take advantage of sophisticated recognition algorithms in image domain. Besides, although 3D faces have the internal pose-invariance characters, incomplete partial faces often bring a lot of difficulties to most existing 3D algorithms.

Sparse Coding. Sparse coding aims to represent each input instance $\mathbf{x} \in \mathbb{R}^d$ using a set of basis vectors $\{\mathbf{b}_j \in \mathbb{R}^d, j = 1, \dots, n\}$ with a sparse coefficient vector $\mathbf{s} \in \mathbb{R}^n$, such that $\mathbf{x} \approx \sum_j \mathbf{b}_j s_j$. Let us denote by matrix $X \in \mathbb{R}^{d \times N}$ for a set of N input data instances, denoted by $B = [B_1, B_2, \dots, B_n] \in \mathbb{R}^{d \times n}$ the basis matrix, and denoted by $S \in \mathbb{R}^{n \times N}$ the sparse coefficient matrix, then the sparse codes problem could be formulated as the following optimization problem [10]:

$$\min_{B, S} \frac{1}{2} \|BS - X\|_F^2 + \lambda \sum_{i,j} \phi(S_{ij}) \quad s.t. \quad \|B_i\|^2 \leq c, \forall i = 1, \dots, n. \quad (1)$$

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, $\phi(\cdot)$ is a sparsity penalty function (a typical choice is an L_1 penalty function, i.e., $\phi(s_j) = \|s_j\|_1$), c is a constant, and λ is a parameter to balance tradeoff between *fitness* and *sparsity*.

In recent years, a variety of sparse coding algorithms have been proposed to improve sparse coding, such as the efficiency and optimization issues [10]. Meanwhile, sparse coding is also widely applied for many applications. For example, sparse coding has been applied to face recognition tasks in literature [15]. Finally, we note that there are also a lot of emerging studies that attempt to improve the performance of sparse coding techniques by different ways [47].

3 The Proposed Pose Invariant 3D Face Recognition

We first present the system architecture of the proposed Pose Invariant Face Recognition scheme. Figure 1(a) shows the system flow of the proposed PIFR solution. We discuss the details of each step below.

Capturing Faces. We install a 3D camera to capture raw 3D facial mesh data and 2D facial images at the same time. The 2D facial images are mainly used by conventional 2D image based face recognition methods as baseline for empirical comparison.

Face Extraction. We have developed some automated tools to detect and extract facial regions from the raw data using the state-of-the-art AAM algorithm [6];

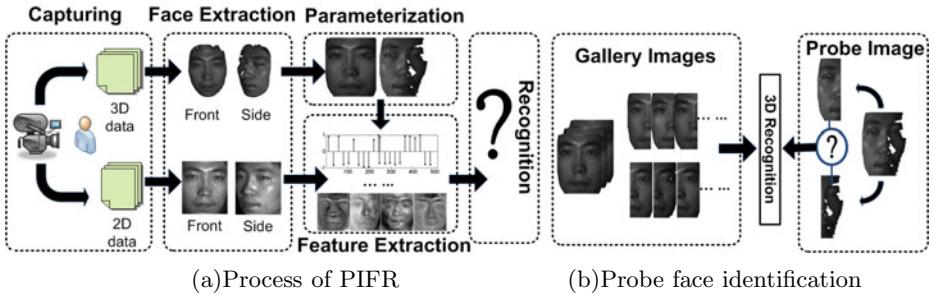


Fig. 1. The system flow of the proposed Pose Invariant Face Recognition scheme

Parameterization. We adopt the state-of-the-art Inverse Curvature Map (ICM) [16] for 3D parameterization, which will be further discussed in Section 4.

Feature Extraction. We propose a novel Locality Preserving Sparse Coding (LPSC) algorithm, which can extract potentially more salient facial features for the recognition tasks. The details will be discussed in Section 5.

Recognition. We simply employ the linear SVM for classification task. We believe it is possible to further boost the performance of our system by kernel based SVM, but the classifier design is out of our focus in this paper.

Recognition of Unseen Faces. The assumptive situation for our recognition task is : there are only front-view 3D mesh faces for each individual in gallery and the probe faces are captured from a very different pose.

In our approach, since we use only one single 3D camera, it is impossible to capture the complete 3D mesh for the side-view faces. However, we are aware a fact that the 3D camera can always capture at least half of the complete human face no matter how the person rotates his/her head (within 90°). Motivated by this practical trick, we propose a half-face recognition approach that automatically finds the complete half-face for prediction, which empirically works effectively, as shown in Figure 1(b). The left-hand side shows the gallery images where each full face is cut to two pieces. During the recognition phase, given an input novel face for prediction, our system automatically generates the complete half-face on-the-fly, and employs it for performing the recognition.

4 Geometry Alignment via 3D Mesh Parametrization

In this section, we present a geometry-based face alignment by applying an effective 3D mesh parameterization technique to face recognition applications. It is important to note that the captured 3D faces may have holes and very different boundaries due to the various poses and occlusions. Since the conventional harmonic map based parameterization highly depends on the boundary, they can hardly be used for the pose-invariant 3D face recognition.

To parameterize the 3D faces, we adopt the inverse curvature map (ICM) [16], which will minimize both the angle distortion and the area distortion by finding the best discrete conformal mapping. The reason is two-fold: firstly, ICM is a boundary free method, thus well-suited for the captured 3D faces with irregular boundaries and even holes. Second, ICM is an intrinsic curvature diffusion approach that only depends on the first fundamental form, i.e., the edge length of the input mesh.

Figure 2 shows a series of 2D faces and parameterized 3D faces for comparison, where the first row shows a series of 2D facial images, the second row shows a series of 3D facial modes, and the last row shows the corresponding aligned facial images by 3D face parameterization, each of which is overlapped with a aligned front-view image as the background image. The alignment results indicate that the geometry-based alignment can effectively align the 3D mesh into a well-posed 2D domain.



Fig. 2. Examples of facial images used in our experiment. The first row shows a series of 2D facial images; the second row shows a series of 3D facial modes; and the third row shows the corresponding aligned facial images by 3D face parameterization, each of which is overlaid with a aligned front-view image as the background image.

5 Locality Preserving Sparse Coding for Facial Images

In this section, we roughly introduce the new proposed sparse coding scheme, named Locality Preserving Sparse Coding (LPSC). The LPSC is designed to address some limitation of the existing sparse coding technique and enhance its performance for face recognition.

Formulation. One key limitation of the existing sparse coding method is that each input vector has been treated equally and independently without exploiting the input data dependency. On the other hand, face images are widely considered to reside on a non-linear submanifold space. This assumption could be used as priori knowledge and explicitly included in sparse coding algorithm.

In order to capture the dependency in input instances, we introduce the following regularizer $g(S, W)$ which measures the inconsistency between the learned

sparse code representation S and the weight matrix W of the input patterns, following the manifold regularization approach [3]:

$$g(S, W) = \frac{1}{2} \sum_{i,j=1}^N W_{ij} \|\mathbf{s}_i - \mathbf{s}_j\|^2 = \text{tr}(SLS^\top) \quad (2)$$

where $\text{tr}(\cdot)$ is the *trace* function, and $L = D - W$, $D = \text{diag}(d_1, \dots, d_N)$ is the degree matrix with the diagonal elements defined as $d_i = \sum_{j=1}^N w_{ij}$.

Using the above regularizer, we can modify the original optimization problem of sparse coding as follows:

$$\min_{S, B} \frac{1}{2} \|BS - X\|_F^2 + \lambda_1 \sum_{i=1}^N \|\mathbf{s}_i\|_1 + \lambda_2 \text{tr}(SLS^\top) \quad \text{s.t. } \|B_i\|^2 \leq c, \forall i = 1, \dots, n. \quad (3)$$

where λ_1 and λ_2 are two regularization parameters.

In order to solve the above optimization problem, we separate the learning process of Locality Preserving Sparse Coding into two optimization tasks: (1) *Coefficient Learning*, i.e., find the solution of S by fixing the dictionary B ; and (2) *Dictionary Learning*, i.e., find the solution of dictionary B by fixing S .

Coefficients Learning. By fixing the dictionary B , the optimization in Eq. (3) reduces to a convex optimization. In our approach, we employ a coordinate descent approach for solving the optimization iteratively. In particular, we iteratively optimize only one of the N coefficient vectors \mathbf{s}_i by leaving the other coefficient vectors intact at one time, and repeat until convergence is arrived.

Specifically, consider iteration t , the solution to $\mathbf{s}_i^{(t)}$ can be found by solving the following optimization:

$$\arg \min_{\mathbf{s}_i^{(t)}} \frac{1}{2} \|B\mathbf{s}_i^{(t)} - \mathbf{x}_i\|_2^2 + \lambda_1 \|\mathbf{s}_i^{(t)}\|_1 + \lambda_2 L_{ii} \|\mathbf{s}_i^{(t)}\|_2^2 + 2\lambda_2 L_i Z^\top \mathbf{s}_i^{(t)} \quad (4)$$

where $L_i^\top \in \mathbb{R}^{N-1}$ is the i -th vector of L by removing the element L_{ii} , and $Z \in \mathbb{R}^{n \times (N-1)}$ is the sub-matrix of $S^{(t)}$ by removing its i -th column vector.

The above optimization is known as a nonsmooth L1 minimization problem. In our approach, we develop an efficient optimization algorithm to solve the above problem by adapting the state-of-the-art non-smooth convex optimization technique proposed in [11], which is able to achieve a fast convergence of $O(1/t^2)$.

Dictionary Learning. Once the coefficient matrix S is found, the other task is to learn the dictionary B given the matrix S . Specifically, when S is given, the optimization of (3) can be reduced to the following:

$$\min_B \|X - BS\|_F^2 \quad \text{s.t. } \|B_i\|^2 \leq c, \forall i = 1, \dots, n. \quad (5)$$

The above optimization is essentially the same as the Dictionary learning task of the regular sparse coding method. In our approach, we adopt the existing dictionary learning algorithm proposed in [10].

Out-of-Sample Coding. Consider a set of training data points $X^{(train)} \in \mathbb{R}^{d \times N}$, by applying the previous algorithm we are able to obtain the optimal dictionary matrix $B^* \in \mathbb{R}^{d \times n}$ and the coefficients matrix $S^* \in \mathbb{R}^{n \times N}$ for the training data. Given an unseen/test data instance $x^{(test)} \in \mathbb{R}^d$, putting $x^{(test)}$ together with the collection of training data points, we can update the Laplacian matrix $L' \in \mathbb{R}^{(N+1) \times (N+1)}$ with $[X^{(train)}, x^{(test)}] \in \mathbb{R}^{n \times (N+1)}$ to capture the dependency between the test instant and training instants.

To find the coefficient vector \hat{s} without modifying the existing coefficients S^* , we can simply solve the following optimization:

$$\arg \min_{\mathbf{s}} \frac{1}{2} \|B^* \mathbf{s} - \mathbf{x}^{(test)}\|_2^2 + \lambda_1 \|\mathbf{s}\|_1 + \lambda_2 L_{(N+1)(N+1)} \|\mathbf{s}\|_2^2 + 2\lambda_2 \mathbf{v}(S^*)^\top \mathbf{s}$$

where $\mathbf{v} = \{v_1, v_2, \dots, v_N\}$ is the $(N+1)$ -th row of L' by removing the last element $L'_{(N+1)(N+1)}$.

6 Experiments

6.1 Experimental Testbed

In our database, we collected totally 10 individuals with different head positions, including the front-view faces and side-view faces with rotation angle varying from 10° to 75° in-depth. Figure 2 already shows some example images of varied poses in our database.

6.2 Evaluation of LPSC for 2D Face Recognition

This section mainly aims to examine the efficacy of the proposed LPSC algorithm. To this purpose, we apply our algorithm to the benchmark face recognition task using the well-known ORL face database and YALE face database.

The ORL database contains 400 facial images from 40 individuals and the YALE face database contains 160 facial images from 15 individuals with different expression and illumination condition. For comparison, we adopt four famous baseline techniques: the Sparse Coding (SC) method [10], the LDA method [2], the LPP method [8] and the PCA method [14].

For the ORL database, we build the training sets by randomly choosing M (2,4,6,8) images for each individual, and putting the rest facial images to form the test set. For the parameter selection, for each data partition, the best feature dimension of LDA, LPP and PCA are chosen by maximizing the classification performance on the training set. For both SC and LPSC methods, we chose the number of basis vector as 1.5 times the input data dimension to balance the sparsity and computation complexity. Similar, the other parameters for SC and LPSC, and the cost parameter C for linear SVM, are found by validating their classification performance on the training set.

Table 1. Result of ORL database

M	LPSC	SC	LDA	LPP	PCA
2	0.7981 ±0.015	0.7926 ±0.014	0.7770 ±0.012	0.7800 ±0.013	0.7134 ±0.009
4	0.9314 ±0.016	0.9176 ±0.021	0.9000 ±0.021	0.9030 ±0.017	0.8800 ±0.009
6	0.9680 ±0.012	0.9527 ±0.015	0.9400 ±0.012	0.9410 ±0.015	0.9225 ±0.012
8	0.9867 ±0.012	0.9589 ±0.020	0.9520 ±0.014	0.9550 ±0.020	0.9300 ±0.008

Table 2. Result of YALE database

M	LPSC	SC	LDA	LPP	PCA
3	0.7006 ±0.018	0.6889 ±0.020	0.6828 ±0.033	0.6889 ±0.043	0.6111 ±0.048
5	0.7963 ±0.038	0.7756 ±0.036	0.7444 ±0.042	0.7502 ±0.035	0.6867 ±0.033
7	0.8478 ±0.023	0.8122 ±0.035	0.7933 ±0.043	0.8011 ±0.043	0.7733 ±0.039
9	0.8711 ±0.038	0.8370 ±0.033	0.8063 ±0.053	0.8133 ±0.054	0.7919 ±0.055

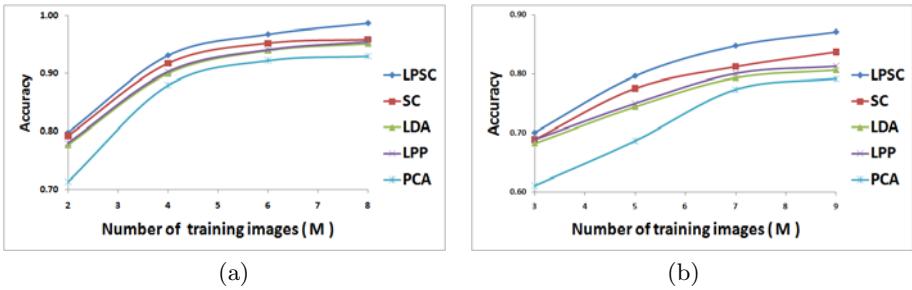


Fig. 3. Experimental results of facial image recognition on the ORL database and YALE database. X-axis value M denotes the number of randomly selected training images from each person. (a) Results of ORL database (b) Results of YALE database.

Figure 3(a) shows the average experimental results over ORL database with varied number of training images for each individual(M). Table 1 illustrates the mean value and the standard deviation in details. For the YALE database, we conduct the similar experimental evaluation with varied $M = 3, 5, 7, 9$. The experiment results of YALE database are shown in Figure 3(b) and Table 2.

Several observations can be drawn from the results. First of all, for both of the two databases the proposed LPSC method achieved the best overall performance among all the compared methods for varied M values. For example, considering the case of $M = 8$ for the ORL database, the average accuracy achieved by LPSC is about 98.67%, which is much higher than the others, including SC (95.89%), LDA(95.20%), LPP(95.50%) and PCA(93.00%). Second, it is evident that the larger the number of training images (M), the better the recognition performance achieved by all the compared methods. Finally, the performance difference between LPSC and SC becomes more significant when increasing the value of M , which indicates that the larger the training size, the more data dependency information can be exploited by LPSC.

As a summary, the above results showed that the proposed LPSC method can learn more effective features for improving the face recognition tasks.

6.3 Evaluation of Pose Invariance Face Recognition : 2D vs. 3D

To evaluate the performance of the proposed pose invariant 3D face recognition system, we compare our solution with regular 2D face recognition in two settings: (1) **Front-Front** recognition task, and (2) **Front-Side** recognition task.

Task I: Front-Front Recognition. For the Front-Front task, our PIFR system is similar to a regular 2D face recognition system because no pose variance should be exploited. Our goal is to evaluate whether the parameterized faces have the same efficiency as the regular front-view face recognition.

In our experiments, we follow the experiment setting in the Section 6.2. The same five algorithms (LPSC, SC, LDA, LPP and PCA) are engaged for comparison following the similar experimental scheme. We perform the recognition task over 2D and 3D front-view images respectively. Table 3 shows the average recognition accuracy and their standard deviations achieved by different algorithms. Figure 4 further illustrates these results.

Two observations can be drawn from these results. First of all, for both 2D faces and 3D faces, the accuracies of the Front-Front recognition tasks are very high. There is no significant difference between 2D and 3D approaches according to statistical *t*-test. Second, similar to the previous observations, LPSC performs the best for both 2D cases and 3D cases among all the compared methods.

As a summary, for the Front-Front task, no significant difference exists between regular face recognition systems and our PIFR system, which shows that our solution is comparable to the regular 2D recognition systems for simple and easy face recognition tasks without pose variations.

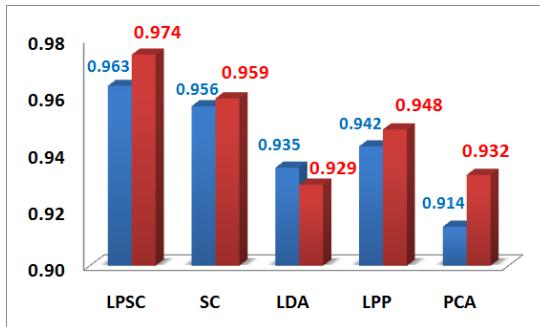


Fig. 4. Comparison of 2D and 3D Front-Front recognition results

Table 3. Comparison of 2D and 3D Front-Front Face Recognition

	2D	3D
LPSC	0.9634 ±0.033	0.9744 ±0.036
SC	0.9562 ±0.033	0.9590 ±0.029
LDA	0.9345 ±0.043	0.9285 ±0.032
LPP	0.9420 ±0.042	0.9480 ±0.037
PCA	0.9137 ±0.033	0.9320 ±0.020

Task II: Front-Side Recognition. This task assumes that there are only a small size of front-view 2D or 3D images in the gallery (5 images for each person). On the other hand, the test faces consist of pose-variant images with large rotation angles varying from 10° to 75°(about 5 to 10 images for each person).

In this experiment, we engage the same five algorithms for both 2D and 3D recognition and found their parameters similar to the former experiments. Moreover, for the 3D face recognition, we train two different models for the left and right half-faces, respectively. Further, our system can automatically detect the complete half-face corresponding to the probe partial 3D faces, so as to choose the corresponding model for recognition.

In order to evaluate pose tolerance ability of our system, we also compare our scheme with the state-of-the-art algorithm (TFA) [12] based on the image space for fairness, because the former five algorithms all utilize the whole images as input instances. The 2D face images are broadly divided into 7 poses ($0^\circ, \pm 10^\circ, \pm 30^\circ$ and $\pm 75^\circ$) for model training and recognition. The further experiment follows the description of the experiment section in [12].

Figure 5 shows the comparison of 2D and 3D Front-Side recognition for $M = 4$. Several observations can be drawn. First of all, for this challenging task, the regular face recognition system performs poorly, with the best accuracy of about 50.00%. Second, the state-of-the-art pose-invariant TFA algorithm seems to be fairly effective for this challenging pose-invariant face recognition task, with an average recognition accuracy of about 80%.

Finally, the proposed PIFR system achieved a much higher accuracy than the TFA algorithm. In particular, the accuracy of our system is about 83% with PCA, which is further boosted to 90% with the proposed LPSC algorithm. Similar to the TFA study [12], we believe that the performance of our system could be further enhanced by introducing local features that are less sensitive to pose rotations.

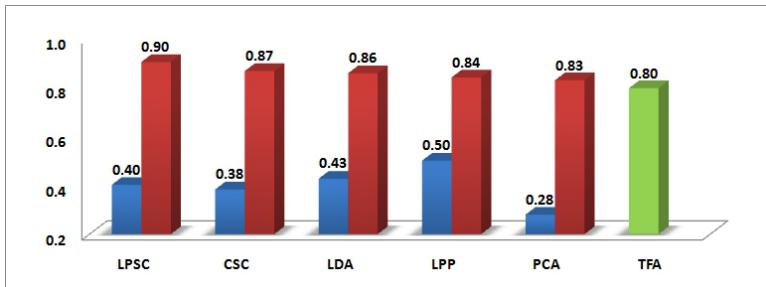


Fig. 5. Comparison of 2D and 3D Front-Side recognition results ($M = 4$). The blur and red bars correspond to 2D and 3D results, respectively, and the green bar represents the result by TFA.

To evaluate the performance of our PIFR system under different m values (from 1 to 4), Figure 6 and Table 4 give more comparison results. The similar observations further confirm the efficacy of LPSC.

As a summary, our empirical results show that the proposed PIFR system can effectively handle the challenging pose variance problem and achieve a comparative result with the state-of-the-art pose-invariance face recognition scheme. By

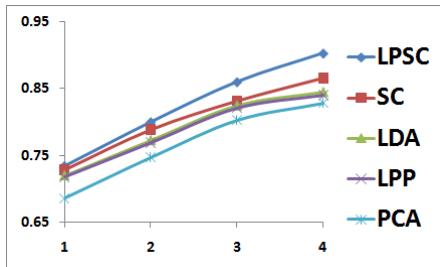


Fig. 6. Results of using different numbers of front face images per individual

Table 4. Results of 3D Front-Side recognition task with different front images

M	LPSC	SC	LDA	LPP	PCA
1	0.7343	0.7286	0.7199	0.6977	0.6857
	±0.020	±0.014	±0.028	±0.018	±0.014
2	0.8000	0.7886	0.7725	0.7589	0.7468
	±0.005	±0.015	±0.016	±0.010	±0.014
3	0.8600	0.8314	0.8242	0.8257	0.8029
	±0.012	±0.006	±0.016	±0.014	±0.016
4	0.9029	0.8657	0.8443	0.8400	0.8286
	±0.006	±0.014	±0.011	±0.020	±0.006

combining the proposed LPSC feature extraction algorithm, its accuracy could be further significantly boosted.

7 Conclusions

This paper proposed a new and effective pose-invariant 3D face recognition scheme equipped with a set of effective parameterization, alignment and feature representation techniques. Our extensive experiments demonstrated the proposed scheme is effective and promising for tackling the PIFR task and achieve better results than the state-of-the-art TFA algorithm. For future work, we will investigate 3D recognition techniques to address the other challenging face recognition tasks, such as the variations of illumination and expression issues, making our 3D face recognition system practical for real-world applications.

Acknowledgement

The work was supported by the Singapore National Research Foundation Interactive Digital Media R&D Program, under research grant NRF2008IDM-IDM004-006.

References

1. Abate, A.F., Nappi, M., Riccio, D., Sabatino, G.: 2d and 3d face recognition a survey. *Pattern Recognition Letters* 28(14), 1885–1906 (2007)
2. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(7), 711–720 (1997)
3. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research* 7, 2399–2434 (2006)
4. Bengio, S., Pereira, F., Singer, Y., Strelow, D.: Group sparse coding. In: Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C.K.I., Culotta, A. (eds.) *Advances in Neural Information Processing Systems*, vol. 22, pp. 82–89 (2009)

5. Bowyer, K.W., Chang, K., Flynn, P.: A survey of approaches to three-dimensional face recognition. *Pattern Recognition* 1, 358–361 (2004)
6. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23(6), 681–685 (2001)
7. Gao, S., Tsang, I.W., Chia, L.-T., Zhao, P.: Local features are not lonely - laplacian sparse coding for image classification. In: *CVPR* (2010)
8. He, X., Yan, S., Hu, Y., Niyogi, P.: Face recognition using laplacianfaces. *IEEE Trans. Pattern Anal. Mach. Intell.* 27(3), 328–340 (2005)
9. Hoi, C.H., Lyu, M.R.: Robust face recognition using minimax probability machine. In: Proc. IEEE Conf. on Multimedia and Expo. (ICME 2004), pp. 1175–1178 (2004)
10. Lee, H., Battle, A., Raina, R., Ng, A.Y.: Efficient sparse coding algorithms. In: *NIPS*, pp. 801–808. NIPS (2007)
11. Nesterov, Y.: Smooth minimization of non-smooth functions. *Mathematical Programming* 103(1), 127–152 (2005)
12. Prince, S., Warrell, J., Elder, J., Felisberti, F.: Tied factor analysis for face recognition across large pose differences. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(6), 970–984 (2008)
13. Tan, X., Chen, S., Zhou, Z.-H., Zhang, F.: Face recognition from a single image per person: A survey. *Pattern Recognition* 39(9), 1725–1745 (2006)
14. Turk, M.A., Pentland, A.P.: Face recognition using eigenfaces. In: *CVPR*, pp. 586–591 (1991)
15. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(2) (2009)
16. Yang, Y.L., Kim, J., Luo, F., Hu, S.M., Gu, X.: Optimal surface parameterization using inverse curvature map. *IEEE Transactions on Visualization and Computer Graphics* 14(5), 1054–1066 (2008)
17. Zhang, X., Gao, Y.: Face recognition across pose: A review. *Pattern Recognition* 42(11), 2876–2896 (2009)
18. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. *ACM Comput. Surv.* 35(4), 399–458 (2003)

Score Following and Retrieval Based on Chroma and Octave Representation

Wei-Ta Chu and Meng-Luen Li

Department of Computer Science and Information Engineering,

National Chung Cheng University,

Chiayi, Taiwan 621

wtchu@cs.ccu.edu.tw, badbadyuniko@hotmail.com

Abstract. With the studies of effective representation of music signals and music scores, i.e. chroma and octave features, this work conducts score following and score retrieval. To complement the shortage of chromagram representation, energy distributions in different octaves are used to describe tone height information. By transforming music signals and scores into sequences of feature vectors, score following is transformed as a sequence matching problem, and is solved by the dynamic time warping (DTW) algorithm. To conduct score retrieval, we modify the backtracking step of DTW to determine multiple partial matchings between the query and a score. Experimental results show the effectiveness of the proposed features and the feasibility of the modified DTW algorithm in score retrieval.

Keywords: Score following, score retrieval, chroma, octave, sequence matching.

1 Introduction

With the rapid popularity of digital music, people usually have huge music collections. People who love music usually tend to follow the corresponding music score when a music piece is played. From another viewpoint, musicians in concerts usually have to manually turn pages of sheet music. For a professional performer, it is a routine for page turning during live performance. However, tuning pages is uncomfortable for a beginner and may distract performance. To address these issues, finding temporal matching between a music signal and the corresponding score is necessary.

We develop a framework as shown in Figure 1 to conduct music-score matching so as to achieve score following. Chroma features and octave features are extracted from both music signals and music scores. Based on the same representation, an approximate sequence matching algorithm is used to match two sequences, and the determined matching represents temporal correspondence between two media. We therefore achieve score following, in which a score bar can be highlighted when its corresponding music segment is being played. With the module that evaluates similarity between a music signal and a music score, we extend this framework to conduct score retrieval, which is similar to query by humming.

Contributions of this work mainly lie on the usage of novel features. We elaborately explore characteristics of music signals, and thus extract chroma and octave

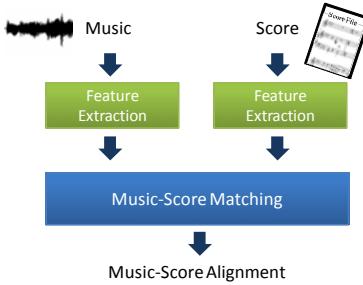


Fig. 1. The framework of music-score alignment

information. In experiments, we verify that the proposed features provide promising performance, as comparing with conventional energy-based and pitch-based features.

The rest of this paper is organized as follows. Section 2 gives brief literature survey. Section 3 describes the features used to describe music signals and scores, in which octave features are newly proposed. Details of score following and score retrieval are described in Section 4. Section 5 provides extensive evaluation results, following by conclusion of this work in Section 6.

2 Related Work

By definition, score following is to align part of a music score to the corresponding portion of a music signal. The approach in [1] uses a two-level hidden Markov model (HMM) to handle score following for polyphonic instruments, such as the piano and guitar. A real-time decoding scheme was designed to return the HMM state sequence corresponding to the alignment between music and score. A Viennese company Qidenus [2] develops a page-turning device, which can be controlled by foots and supports musicians to avoid inconvenient manual page-turning. In [3], this page-turner device is utilized to automate page-turning of sheet music. They first convert sheet music into MIDI files, and then extract audio features from them. The positions where the pages should be turned are labeled manually. From the lively performed music signal, audio features are also extracted so that feature sequences from music and score are aligned by the dynamic time warping algorithm. Recently, the importance of score following has been widely recognized, and more and more researchers are involved. Therefore, the annual evaluation campaign “Music Information Retrieval Exchange” [4] starts the task of score following from year 2006.

3 Feature Extraction

3.1 Chroma and Octave Features

Shepard advocated that human’s pitch perception is like a helix, in which the vertical and angular dimensions represent *tone height* and *chroma*, respectively [5]. Music notes in an octave can be classified into twelve pitch families, i.e. chroma. In this helix, as the pitch increases (e.g. from C0 to C1), it looks like moving along the helix,

passing through other pitch families chromatically, and finally going back to the same family (C) that is one octave higher than the initial point. It means that two different music notes could be grouped into the same pitch family if their corresponding frequencies have some relationship. For example, in Figure 2(a), the first and the eighth music notes are both mapped to the chroma C. With this idea, every audio frame is transformed into a 12-dimensional chromagram, in which each dimension represents the accumulated energy of a chroma, respectively.

Chromagram only conveys information of pitch families. Therefore, music sequences with the same evolution of chroma changes may have the same representation. In Figure 2, although music notes of these three segments are different, their chromagrams are the same. In this work, we propose an 8-dimensional octavegram to address this issue. An octave is composed of 12 semi-tones (i.e. C, C#, D, #D, etc.), and the frequency domain can be divided into eight octaves. Every audio frame can be transformed into an 8-dimensional octavegram, in which each dimension represents energy corresponding to one octave, respectively. With this representation, three segments in Figures 2(a), (b), and (c) can be discriminated.

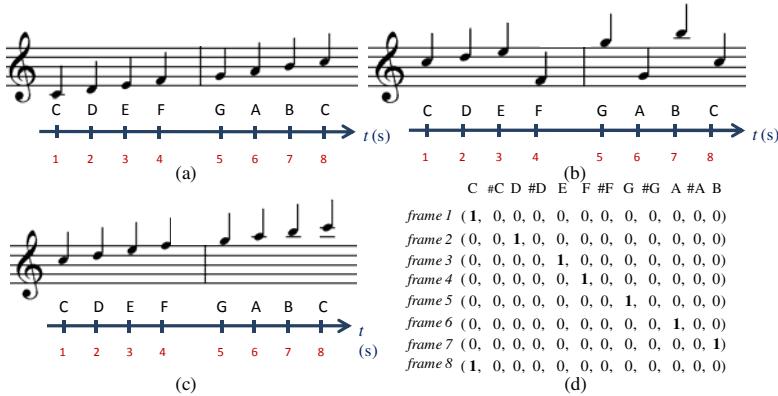


Fig. 2. (a)(b)(c) Three different music segments; (d) The chromagram corresponding to all (a), (b), and (c)

3.2 Feature Extraction from MIDI

To efficiently represent music scores, which are represented by MIDI format in this work, we parse bit streams of MIDI files to extract music scores, and then compute corresponding chroma features and octave features. A tuple consisting of two components is used to describe the information of music notes:

$$(N, duration), \quad (1)$$

where $N = \{n_1, n_2, \dots, n_K\}$ is a set of music notes that strike simultaneously, and n_i denotes the MIDI node id.

Assume that a MIDI file has M tracks, and the i th track has j_i tuples. This MIDI file is then represented as M sequences (S_1, S_2, \dots, S_M) , in which $S_i = \{(N_{i1}, t_{i1}), \dots, (N_{ij_i}, t_{ij_i})\}$. Figure 3 shows a MIDI track consisting of six tuples. The first tuple, for example, consists of four music notes (48, 52, 55, 60) with duration 0.4 seconds.

To calculate chroma value of a music note, an MIDI node is converted into a chroma by

$$c = \begin{cases} (n_k \% 12) + 1, & \text{for } n_k \geq 0, \\ -1, & \text{for } n_k < 0. \end{cases} \quad (2)$$

We divide the MIDI-based score into non-overlapping 0.2-second segments, called score frames. For each score frame, we count the number of chroma occurring in this segment, and describe this segment as a 12-dim chroma vector. A music score is therefore transformed into a sequence of chroma vectors.

Figure 4 illustrates the process for converting Figure 3 into chroma vectors. The first tuple ($\{48, 52, 55, 60\}, 0.4$) is first converted into chroma representation ($\{C, E, G, C\}, 0.4$). At the first score frame (0~0.2 second), the chroma C occurs twice since both notes 48 and 60 map to C. The normalized value of the k th chroma bin of the chroma vector $\tilde{\mathbf{p}}_f^c$ at the score frame f can then be calculated as

$$\tilde{\mathbf{p}}_{f,k}^c = \frac{\mathbf{p}_{f,k}^c}{\sum_{j=1}^{12} \mathbf{p}_{f,j}^c}. \quad (3)$$

To calculate octave features from MIDI, an MIDI node is converted into an octave number by

$$o = \begin{cases} \lfloor n_k / 12 \rfloor + 1, & \text{for } n_k \geq 0, \\ -1, & \text{for } n_k = -1. \end{cases} \quad (4)$$

With the same idea of calculating chroma features, we divide the MIDI-based score into non-overlapping 0.2-second score frame. For each score frame, we count the number of octaves in which some music notes strike, and describe this segment as an 8-dim octave vector, since there are totally eight octaves in the frequency domain. A music score is therefore transformed into a sequence of octave vectors.

The normalized value of the k th octave bin of the octave vector $\tilde{\mathbf{p}}_f^o$ at the score frame f can then be calculated as

$$\tilde{\mathbf{p}}_{f,k}^o = \frac{\mathbf{p}_{f,k}^o}{\sum_{j=1}^8 \mathbf{p}_{f,j}^o}. \quad (5)$$

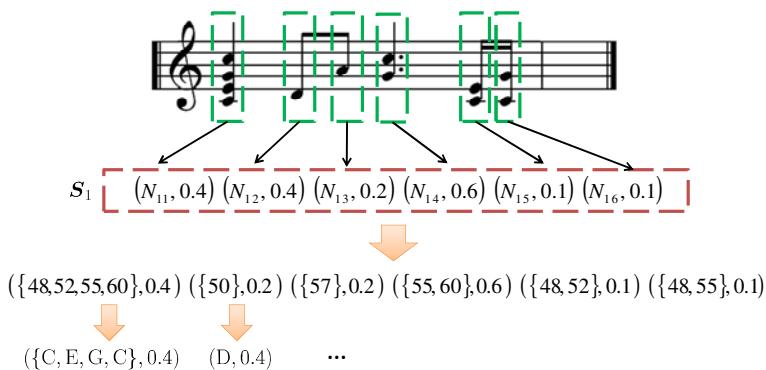


Fig. 3. An example of track sequence

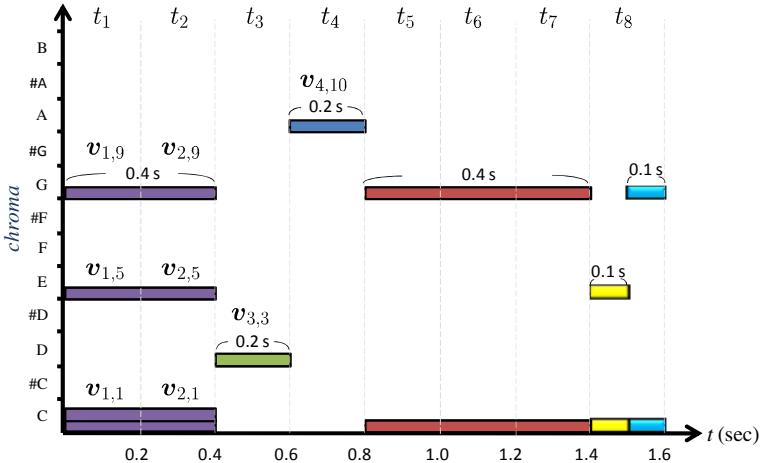


Fig. 4. An example for calculating energies in different bins

3.3 Feature Extraction from Audio

To efficiently describe audio signals, audio data are first transformed into the frequency domain, and then energy of each frequency bin is calculated. In experiments of this work, we collect music pieces from [6] because this website contains music performed by different instruments and the corresponding music scores stored in MIDI format. Each music piece is interpreted by a player rather than computer-synthesized, and there may be noise in music. Therefore, we propose a pre-emphasis method to strengthen energy distribution in each frame.

First, the frequency domain is divided into eight regions, and each region contains twelve frequency bins. Let $E_{f,k}$ be the energy of the k th frequency bin at frame f , and $E_{f,r}$ be the accumulated energy of the r th region, where $r \in [1, 8]$. The average energy of the r th region at frame f is calculated as $\text{Avg}E_{f,r} = E_{f,r}/12$. With the previous definitions, the pre-emphasis process is formulated as

$$\hat{E}_{f,k} = \begin{cases} w_e \times E_{f,k}, & \text{for } E_{f,k \in S_r} \geq \text{Avg}E_{f,r}, \\ E_{f,k}, & \text{otherwise,} \end{cases} \quad (6)$$

where $w_e \geq 1$ indicates the *energy weight*, and S_r is a set of frequency bins that corresponds to the r th region.

After pre-emphasis, we calculate the corresponding chromagram. The accumulated energy $\tilde{E}_{f,h}$ of each pitch family h is calculated as

$$\tilde{E}_{f,h} = \sum_{k \in S_h} \hat{E}_{f,k}, \quad (7)$$

where S_h is a set of frequency bins that corresponds to the chroma family h . The normalized value of the k th chroma bin of the chroma vector $\tilde{\mathbf{q}}_f^c$ at the music frame f can then be calculated as

$$\tilde{\mathbf{q}}_{f,k}^c = \frac{\tilde{E}_{f,k}}{\sum_{j=1}^{12} \tilde{E}_{f,j}}. \quad (8)$$

Examples of chromagram can be seen in Figure 2(d).

To generate an 8-dim octave vector, we calculate energy of each frequency bin, and the accumulated energy $\tilde{E}_{f,g}$ of each element of an octave vector could be calculated as

$$\tilde{E}_{f,g} = \sum_{k \in S_g} \hat{E}_{f,k}, \quad (9)$$

where S_g is a set of frequency bins that corresponds to the octave family g . The normalized value of the k th octave bin of the octave vector $\tilde{\mathbf{q}}_f^o$ at the music frame f can then be calculated as

$$\tilde{\mathbf{q}}_{f,k}^o = \frac{\tilde{E}_{f,k}}{\sum_{j=1}^8 \tilde{E}_{f,j}}. \quad (10)$$

4 Score Following and Score Retrieval

4.1 Music-Score Matching

We have transformed both music signals and music scores into sequences of feature vectors, in the representation of chroma and octave. The problem of music-score matching is thus transformed into a sequence matching problem. Assume that there are m music frames and n score frames in these two sequences. Let $\mathbf{Q}^c = (\mathbf{q}_1^c, \mathbf{q}_2^c, \dots, \mathbf{q}_m^c)$ and $\mathbf{Q}^o = (\mathbf{q}_1^o, \mathbf{q}_2^o, \dots, \mathbf{q}_m^o)$ respectively denote the sequences of chroma vectors and octave vectors of a music signal Q . Let $\mathbf{P}^c = (\mathbf{p}_1^c, \mathbf{p}_2^c, \dots, \mathbf{p}_n^c)$ and $\mathbf{P}^o = (\mathbf{p}_1^o, \mathbf{p}_2^o, \dots, \mathbf{p}_n^o)$ respectively denote the sequences of chroma vectors and octave vectors of the score corresponding to the music signal Q . Note that we simplify the notations $\tilde{\mathbf{q}}_i^c, \tilde{\mathbf{q}}_i^o, \tilde{\mathbf{p}}_i^c, \tilde{\mathbf{p}}_i^o$ as $\mathbf{q}_i^c, \mathbf{q}_i^o, \mathbf{p}_i^c, \mathbf{p}_i^o$ to denote normalized vectors. We then respectively construct a cost matrix based on chroma features and octave features. The (i, j) -th entry of the chroma-based cost matrix $M_{i,j}^c$ is calculated as

$$M_{i,j}^c = d(\mathbf{q}_i^c, \mathbf{p}_j^c) = \sqrt{\sum_{\ell=1}^{12} (\mathbf{q}_{i,\ell}^c - \mathbf{p}_{j,\ell}^c)^2}. \quad (11)$$

Similarly, the (i, j) -th entry of the octave-based cost matrix $M_{i,j}^o$ is calculated as

$$M_{i,j}^o = d(\mathbf{q}_i^o, \mathbf{p}_j^o) = \sqrt{\sum_{\ell=1}^8 (\mathbf{q}_{i,\ell}^o - \mathbf{p}_{j,\ell}^o)^2}. \quad (12)$$

We combine two cost matrices by

$$M_{i,j}^b = w_c M_{i,j}^c + (1 - w_c) M_{i,j}^o, \quad (13)$$

where w_c indicates the weight controlling the relative impact of chroma and octave information.

Based on the combined cost matrix, finding the optimal matching between a music signal and its corresponding music score can be solved by finding the alignment with the lowest cost between two feature sequences. This problem is well known to be solved by a dynamic time warping algorithm. Figure 5 shows two examples of alignment results, in which brighter pixels mean smaller costs. The left side shows that the music signal is played strictly according to the score, and the optimal alignment lies on the main diagonal of this cost matrix. The right side shows a matching with slight variations to the main diagonal, because the music signal is played in varied speeds and may be disturbed by noises.

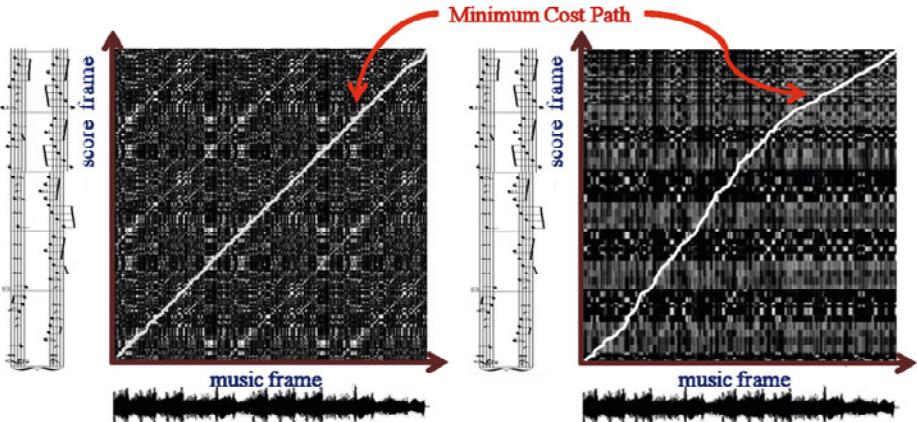


Fig. 5. Two alignment results obtained by the dynamic time warping algorithm

4.2 Score Retrieval

We conduct score retrieval by the same way as music-score matching. Four steps are included in our score retrieval system.

- Feature extraction and cost matrix calculation

The query music piece and score files are converted into sequences of feature vectors. We then calculate a cost matrix between the query music piece and every score in the database, respectively.

- Finding optimal alignment

The dynamic time warping algorithm is used to find an alignment with the minimal cost, while a different implementation is used in the backtracking step. In the classical DTW algorithm, the start point of backtracking is limited at (m, n) , where m and n denote the lengths of two sequences, respectively. This setting is not suitable for score retrieval since the query music piece often simply corresponds to part of a score. The globally optimal alignment between a music score and a short query piece may not give the best solution. Moreover, there may be several similar portions in a score, i.e. the query piece may correspond to multiple segments of a score. Therefore, we do not limit the number of the best alignment to be found.

Instead of always backtracking from the (m, n) -th entry of the cost matrix, we sequentially conduct backtracking from $(m, n), (m, n - 1), \dots, (m, m + 1)$, assuming that the query piece is always shorter than a music signal in the database ($m < n$). An optimal alignment (a path) can be determined from each backtracking case, and we can calculate the average cost of each path based on the entries where the path goes through. Figure 6 shows examples of alignment results, in which P_k , $k = m + 1, m + 2, \dots, n$, denotes a path corresponding to an alignment result of a backtracking case. To eliminate matching noises, a path is filtered out when its average cost is larger than a predefined cost threshold ϵ . The path P_{174} in Figure 6 is discarded, for example.

- Path clustering and filtering

After filtering, several neighboring paths may perceptually correspond to the same alignment, i.e. they nearly indicate the same correspondence between the score and the query. To avoid redundancy, we cluster paths that are obtained by backtracking from neighboring start points. For example, in Figure 6, the remaining paths (say $P_{100}, P_{101}, P_{212}, P_{213}, P_{214}$, and P_{215}) are grouped into two clusters, i.e. $\{P_{100}, P_{101}\}$ and $\{P_{212}, P_{213}, P_{214}, P_{215}\}$. From each cluster, we only choose the path of the smallest average cost to be the representative of a cluster. Recall that each representative path indicates a correspondence between the query and a score. Given a query piece, alignments (representative paths) between it and each score are first found. All paths are then ranked by their average costs, and are returned as the ranked retrieval results.

5 Experiments

Each piece of music used in this work is polyphonic. The standard MIDI file (.mid) is used to store music scores. Three datasets are used in this work. Dataset1 consists of 67 music files and 67 corresponding scores (MIDI files), which are collected from [6]. Each music piece is recorded from a real performance, which proceeds according to a corresponding score. Music pieces have durations ranging from 27.1 to 191.5 seconds. The sampling rate is 44.1kHz and each sample is represented by 16 bits. The amounts of score bars of scores range from 9 to 92. To evaluate performance of music-score matching, we listen to all music files and manually identify how a segment of a music signal corresponds to a score bar in the corresponding music score. There are also 67 music files and 67 corresponding score files (MIDI files) in Dataset2. The score files are same as that in Dataset1, while the music files in Dataset2 are converted from MIDI files by a MIDI synthesizer. Dataset3 is generated for evaluating the score retrieval system. It consists of 1491 music queries, 67 corresponding score files (MIDI files), and 133 irrelevant score files. Each query piece is generated by randomly selecting a segment from a music signal in Dataset1.

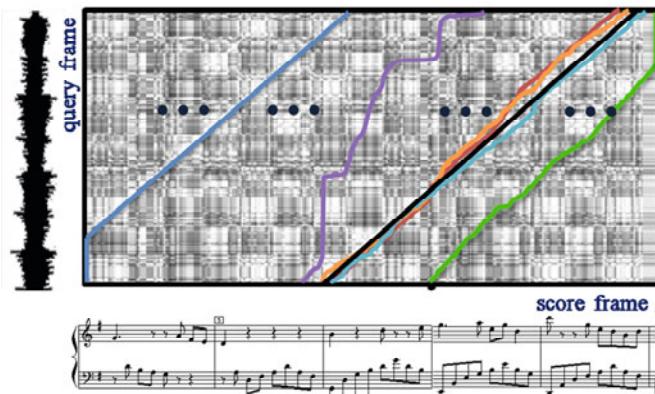


Fig. 6. Various alignment results between the query piece and a music score

5.1 Performance of Music-Score Matching

For two sequences Q and P , the point (i, j) in the alignment path denotes that the music frame q_i is similar to the score frame p_j . In the backtracking path, a point is claimed as a correct matching if it is located in the right score bar. This evaluation scheme tolerates slight skews of alignment results, which are hardly perceived by humans. Accuracy of music-score matching is therefore defined as:

$$\text{accuracy} = \frac{\text{number of correct points}}{\text{number of points in the path}} \times 100\%. \quad (14)$$

Based on Dataset1 and Dataset2, we conduct music-score matching with different cost weights w_c , which controls the importance of chroma vectors, and with different emphasis weights w_e , which controls the degree of pre-emphasis. The value $w_c = 1$ means only the chroma representations is used. The value $w_e = 1$ means pre-emphasis is not applied.

Figure 7 shows accuracy values for Dataset1 and Dataset2, under different weighting settings. We find that an appropriate combination of w_c and w_e provides better music-score matching, while the best settings for two datasets are different. However, adding the pre-emphasis process is not suitable for all the cases. Comparing the results of $w_e = 1$ with that of $w_e = 10$, we observe that the pre-emphasis process improves performance in Dataset1 but not in Dataset2. It means that the pre-emphasis process provides more impacts on music in real performance.

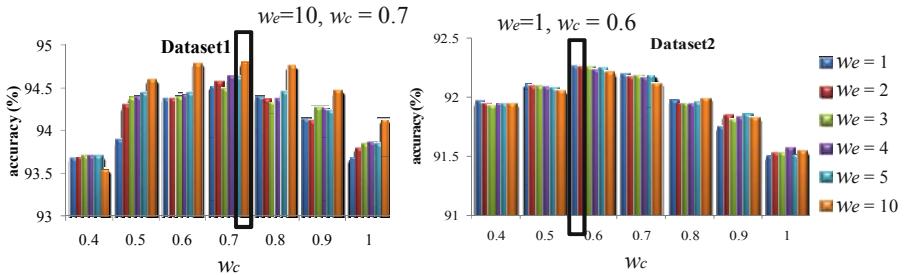


Fig. 7. Accuracies of music-score matching for Dataset1 and Dataset2

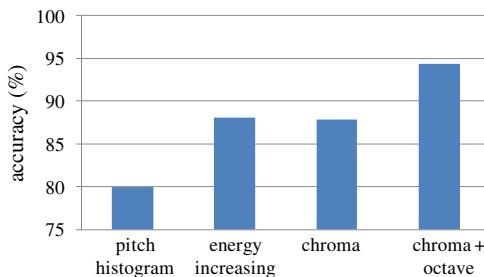


Fig. 8. Performance of music-score matching based on different features

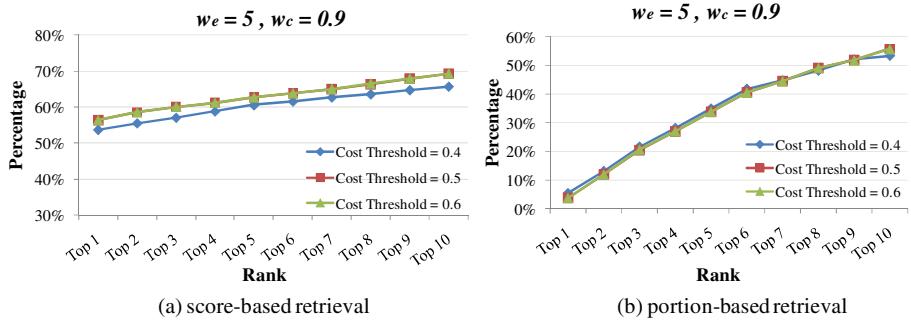


Fig. 9. Performance of score-based retrieval and portion-based retrieval

To further verify the superiority of our proposed features, we compare performance of music-score matching based on four features: 1) pitch histogram [7]; 2) energy increasing feature [8]; 3) chroma [5]; 4) chroma+octave (ours). Figure 8 shows that our approach presents the highest accuracy. Although satisfactory accuracy can be achieved (more than 85%) based on energy increasing features, the high-dimensional representation increases computation cost.

5.2 Performance of Score Retrieval

Two kinds of retrieval are evaluated: 1) *Score-Based Retrieval* and 2) *Portion-Based Retrieval*. In score-based retrieval, since a score file may contain several similar portions, retrieving a score file that really contains a portion similar to the query is regarded as a correct result. In portion-based retrieval, not only the corresponding score should be retrieved, but also the corresponding segments in this score should be correctly indicated.

Dataset3 is adopted in this experiment. For two kinds of retrievals, the results are respectively presented as ranked lists. Figure 9 shows the percentage of queries that retrieve the correct results from the top k returned results, with the best parameter setting ($w_e = 5, w_c = 0.9$). From these results, we found that chroma features play a more important role ($w_c = 0.9$) in music-score matching, no matter in score-based retrieval or portion-based retrieval. Another observation is that the cost threshold ϵ used for filtering out noisy paths doesn't matter a lot, which shows the reliability of the modified DTW algorithm.

6 Conclusion

We have presented score following and score retrieval based on a newly proposed feature and a modified DTW algorithm. In addition to utilize chroma features that are more appropriate to describe music content, we further propose an octave-based feature with a pre-emphasis process to describe energy distributions in different octaves. Score following is transformed into the problem of finding optimal correspondence between two feature sequences, and thus a conventional DTW algorithm is adopted. We modify the backtracking steps of DTW, and extend the idea

to conduct score retrieval. Experimental results show superiority of the proposed feature and feasibility of the proposed pre-emphasis method. In the future, more elaborate features would be investigated to improve matching performance.

Acknowledgement

The work was partially supported by the National Science Council of Taiwan, Republic of China under research contract NSC 99-2221-E-194-036 and NSC 98-2221-E-194-056.

References

1. Schwarz, D., Orio, N., Schnell, N.: Robust Polyphonic MIDI Score Following with Hidden Markov Models. In: Proceedings of the International Computer Music Conference, pp. 129–132 (2004)
2. Qidenus Technologies, <http://www.qidenus.com>
3. Cont, A., Schwarz, D., Schnell, N., Raphael, C.: Evaluation of Real-Time Audio-to-Score Alignment. In: Proceedings of the International Society for Music Information Retrieval, pp. 315–316 (2007)
4. The Music Information Retrieval Exchange (MIREX),
http://www.music-ir.org/mirex/wiki/MIREX_HOME
5. Shepard, R.N.: Circularity in Judgements of Relative Pitch. Journal of the Acoustical Society of America 36, 2346–2353 (1964)
6. Free-Scores.com, <http://www.free-scores.com/>
7. Tzanetakis, G., Ermolinskyi, A., Cook, P.: Pitch Histograms in Audio and Symbolic Music Information Retrieval. In: Proceedings of the International Conference of Music Information Retrieval, pp. 31–38 (2002)
8. Arzt, A.: Score Following with Dynamic Time Warping: An Automatic Page-Turner. Master's Thesis, University of Technology, Vienna (2008)

Incremental Multiple Classifier Active Learning for Concept Indexing in Images and Videos

Bahjat Safadi, Yubing Tong, and Georges Quénot

Laboratoire d'Informatique de Grenoble

BP 53 - 38041 Grenoble Cedex 9, France

{Bahjat.Safadi,Yubing.Tong,Georges.Quenot}@imag.fr

Abstract. Active learning with multiple classifiers has shown good performance for concept indexing in images or video shots in the case of highly imbalanced data. It involves however a large number of computations. In this paper, we propose a new incremental active learning algorithm based on multiple SVM for image and video annotation. The experimental result show that the best performance (MAP) is reached when 15-30% of the corpus is annotated and the new method can achieve almost the same precision while saving 50 to 63% of the computation time.

Keywords: Multimedia Indexing, Machine Learning, Active Learning and Incremental Learning.

1 Introduction

Supervised learning consists in training a system from sets of positive and negative examples. The learning system may be composed of various types of feature extractors, classifiers and fusion modules. The performance of the systems depends a lot upon the implementation choices and details but it also strongly depends upon the size and quality of the training examples. While it is quite easy and inexpensive to get large amounts of raw data, it is usually very expensive to have them annotated because it involves human intervention for the judging of the "ground truth". While the volume of data that can be manually annotated is limited due to the cost of manual intervention, there remains the possibility to select the data samples that will be annotated so that their annotation is as useful as possible (III). *Active Learning* is a special case of machine learning which has been used to improve query performance in image retrieval systems. The objective of Active learning is to maximize the expected information from the query as a result of user feedback in order to minimize the total number needed for the search. This can be summarized as following, from relevance feedback, a user subjectively labels the retrieved images as Positive or Negative, and these labelled images are used to train a classifier that performs a bi-class classification on the image database. Those images with the higher scores or probability values, with respect to the positive image class, are retrieved as the most informative samples to be labelled by the user. However, in very large databases, the learning performance is often restricted due to a very small number of available labelled samples to a given concept, because labelling concepts in too many images or videos is a very hard task, and a user is unwilling to

label too many retrieved images for relevance feedback. Classifiers (such as Support Vector Machines) based active learning has been proposed [3][9][12][13][16] to handle this problem by maximizing the learning efficiency while minimizing the required number of labelled image samples for training process.

Recently, some researches like in [12][17][18] have shown the effectiveness of using the multi-learners approach to handle the problem of the imbalances between the major and minor classes in the very large scale databases. This problem is very common in concept indexing in images and videos since the most target concepts are very sparse. For instances, their average frequency in the TRECVID evaluation campaigns [15] for example, is less than 1%. This imbalance is a serious problem for classical supervised learning methods. An alternative approach to solve this problem is to sub-sampling the major class [4] (the negative samples); this sub-sampling can be done by considering all the positive samples in the training set and selecting randomly a comparable number of negative to positive samples. This sub-sampling might leads to loss of information, due to the fact that it ignores a lot of information from the non chosen negative samples, hence the multi-learner has the ability to balance the loss of information related to this sub-sampling by making several selections on this class and fusing the outputs of different classifiers built from these subsets. In [12], our previous work, we showed that combining between the Active Learning with multi learners approaches significantly increases the effectiveness of the Active Learning, but it makes it very slower comparing to a mono-learner approach. This makes a very big challenge in the task of automatically image annotation, which mostly is directed by learning from user's feedback.

Since during the iteration of active learning and multi-learners (here multiple SVMs are used), new labelled samples will always be added to the training set for next iteration, each iteration involves previous training information and the new untrained samples. The calculation time will be saved if we can re-use the previous information and learn the incremental information derived from new samples. So it is natural to adopt incremental learning for this case.

Some incremental learning focus on how to choose the informative samples data from all the incremental one or retire some samples from previous set [14][20][21]. Those methods needs to check KKT conditions of SVM quadratic optimization problem for every sample which also means much calculation. [6][7] also consider the calculation problem in active learning and multi-learner. An early stopping method is proposed to achieve faster convergence of active learning by counting the number of support vectors derived from previous training in [5]. If the number of support vectors stabilizes, it means that all possible SVs have been selected by active learning method. This method may lose some useful information since the number of SVs may still change after several stable values and the stability of SVs is not clearly defined. Our previous multi-learner active learning research also shows that for every learner in iteration only a part of the samples used for training [12][13]. But many classifiers are still needed to train during iteration. Although some samples in previous step have been well trained but they are never used in the following step. Some researchers have tried some incremental methods used in SVM training. [11] proposed an incremental learning of SVM, that The support vectors from previous training set will be used involved in the new SVM optimization problem with different weights on them. This method can work for

balanced data. But for our extremely imbalanced data, the weight is so rough that the final hyperplane deviated much from the ideal one. Furthermore this method also needs to train the previous support vector, then no time can be saved. In [19] authors proposed an incremental learning of SVM by classifier combining. Multiple SVM are used and each output posterior probability information. For the example incremental learning, the training set can be divided into several learning sequence or incremental batch. Cross validation was made on every batch and classifier is also trained, then the output of each classifier predicting on testing sample can be combined to get the average posterior probability. Here every batch works independently without using the information from previous training.

In this paper, we overcome the problem of processing time for the Active learning with multiple-learners by proposing a robust Incremental algorithm for the Active learning based multiple-SVMs and we show that we can save about 50-63% of the processing time (depending on the used descriptor and the negative to positive ratio), while the system performance was not significantly changed in all cases; our experiments were conducted on the TRECVID 2007 and 2008 collections.

The outline of the paper continues as follows: the combination between the multiple classifier and active learning approach is discussed in section 2. We present our new incremental learning algorithm in introduced in section 3; section 4 describes the experimental results including the description of the used data collection and the descriptors, while Section 5 presents concluding remarks.

2 Active Learning with Multiple Classifiers

Active learning has been adopted to solve problems related to unlabelled training data. For imbalanced data, many papers have shown the possibility to balance the loss of information related to sub-sampling of the negative class by making several selections on this class set and fusing the outputs of different classifiers built from these subsets. This leads to what we call the Multi-learners approach. The active learning algorithm with multiple classifiers is detailed in Algorithm 1 which is a classical active learning algorithm in which we have replaced the single classifier by a set of elementary classifiers. For implementation purposes, the elementary learning algorithm A is split into two parts: Train and Predict. A global parameter, mono-learner, can force the classical active learning mode with a single classifier. At each iteration i , the development set S is split into two parts: L_i , labeled samples and U_i , unlabelled samples. A global parameter f_{pos} defines the ratio between the negative and positive samples in all learners and for all iterations. This defines the number of negative samples for each learner at iteration i . In the multi-learner approach, the number of learners is computed so that each negative sample appears in average a given number of times (usually once) in the different subsets T_j . The T_j contains all positive samples and a randomly chosen subset of negative samples. Classifiers C_j are then trained on the T_j with associated labels and applied to U_i the unlabelled set for the selection of the next samples to annotate. Predictions from the elementary classifiers are then merged in both cases for producing a single prediction score per sample. The predictions on the U_i set are used by Q the selection (or querying) function to produce a sorted list of the next samples to be annotated. From the top of this list, a \tilde{x} set is selected for annotation. The \tilde{x} set is then added

Algorithm 1. Multiple Classifier Active Learning Algorithm

S: all data samples.
 L_i, U_i : labelled and unlabelled subsets of S .
 A =(Train, Predict): the elementary learning algorithm.
 Q : the selection (or querying) function.
 $nl(k)$: number of learners at iteration k .
Initialize L_i (e.g. 10 positives & 20 negatives).
while $S \setminus L_i \neq \emptyset$ **do**
 if mono-learner **then**
 $nl(k) = 1$
 else
 $nl(k) = \text{Calculate the number of Learners}$
 end if
 for all $j \in [1..nl(k)]$ **do**
 Select subset T_j from L_i for training
 $C_j \leftarrow \text{Train}(T_j)$
 $P_{un}^j \leftarrow \text{Predict}(U_i, C_j)$
 end for
 $P_{un} \leftarrow \text{Fuse}(P_{un}^j)$
 Apply Q on P_{un} and select $\tilde{x} \in U_i$ samples.
 $\tilde{y} = \text{Label } \tilde{x}$
 $L_{i+1} \leftarrow L_i \cup (\tilde{x}; \tilde{y})$
 $U_{i+1} \leftarrow U_i \setminus \tilde{x}$
end while

with the associated set of labels \tilde{y} to the L_i set to produce the L_{i+1} set and it is also removed from the U_i set to produce the U_{i+1} set. The global algorithm is determined by the A =(Train, Predict) elementary learning algorithm (e.g. SVM) and by Q the selection (or querying) function implementing the active learning strategy (e.g. relevance or uncertainty sampling). It is also determined by some global parameters like the ratio between the number of negative and positive samples, by the cold start problem, by the fusion function used to fuse the outputs of the classifiers and by the way we choose the number of new samples to be integrated at each iteration. For our evaluation experiments we show the system performance by calculating the Mean average precision on the test set at each step.

3 The Proposed Incremental Method

As described in section 2, algorithm 1 can be used to handle the class imbalance problem. Even though it gave good results [12][13][17], it is still very slow because at each iteration it generates a lot of learners when the dataset is highly imbalanced. In this section we propose an incremental method to reduce the number of learners that need to be trained at each iteration. Let $nl(k)$ be the number of learners needed at step k , $nm[k]$ to be the minimum number of learners to be changed at step k . At each iteration the actual number of learners which should be trained is equal to $nl(k) - (nl(k-1) - mn[k])$. After training these learners we merge their results with the results obtained from

$nl(k - 1) - mn[k]$ learners from the previous steps, so that at each iteration we keep the $nl(k)$ of this iteration but we only train part of them as shown in table 1 where rm and add indicate respectively the number of learners should be removed and added at each iteration.

Table 1. The conditions of removing and adding learners

At step k :
$nl(k)$: number of learners at step k
$nl(k - 1)$: learners trained from the previous steps
$nm[k]$: minimum number of learners to be changed
The number of learners to be removed from the previous step:
<i>if</i> ($nl(k) \geq nl(k - 1)$) $rm = nm$
<i>if</i> ($nl(k) < nl(k - 1)$) $rm = nl(k - 1) - nl(k) + nm$
The number of learners to be added by:
<i>if</i> ($nl(k) \leq nl(k - 1)$) $add = nm$
<i>if</i> ($nl(k) > nl(k - 1)$) $add = nl(k) - nl(k - 1) + nm$

In figure 1 we show one step of how the proposal algorithm works. At each iteration the algorithm should remove the learners with the minimum number of positive samples (normally the learners taken from the oldest iterations), and train some new learners, each learner should be trained on a subset that consists of all the positive samples and of a comparable number of negative samples randomly selected from the training set. At the end it applies the fusion function on the results obtained from the considered learners to give the final score of each unlabelled sample.

In our experiments, we fixed the minimum and maximum values of nm to be $\{1, 10\}$ respectively, and $nm[k] = 20\%nl(k)$.

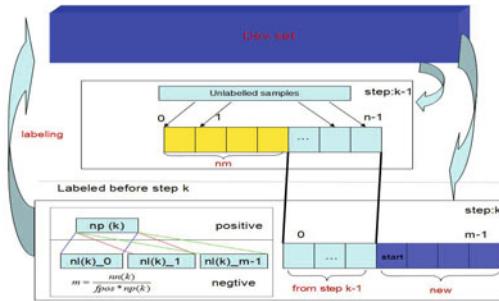


Fig. 1. The framework of the proposed incremental method

4 Experiments

We have evaluated the Multiple Classifiers Active Learning and the proposed incremental methods in a variety of contexts. It has been applied using four types of image descriptors using the SVM with RBF kernel as classifier and the relevance sampling

strategy for active learning. We also used the harmonic Mean function to fuse the results of the multi learners. The cold start problem was not really explored; a random set of 10 positive and 20 negative samples was used. The global parameters like the f_{pos} ratio were taken from our previous work [13]. For the number of samples to be added at each iteration, we chose a variable step size since we observed in previous experiments that having small steps in the beginning of the active learning process is better for the speed of performance improvement. In practice, we used a geometric scale with 40 steps. The evaluations were conducted using the TRECVID 2007 and 2008 test collections and protocols.

4.1 TRECVID 2007 and 2008 Collections

The evaluation was conducted on TRECVID-2008 concepts annotated on the TRECVID 2007 and 2008 collections where 20 concepts were evaluated. The training and evaluation were done respectively on the development and test sets of the two collections, TRECVID 2007 collection contains 21532 video shots as a training set and 22084 shots as test set, while TREC2008 contains 43616 video shots as a training set and 35766 shots as a test set. In the two collections, the training sets are fully annotated and nothing remains to be annotated which makes the use of the Active learning not relevant, but such large fully annotated sets constitute opportunities to simulate, evaluate and compare strategies and methods in active learning without the need of involving a user, as the simulated active learning [2]. In our experiments active learning methods are executed as if very few annotations are available in the training set. Then, each time a human annotation is needed, the corresponding subset of the full annotation is made available to the active learner.

4.2 Image Representation

Concepts and Images can be represented by their vector descriptors or features. Many descriptors could be used to represent a specific concept in an image, finding the best descriptor to represent a concept in an image is still a big challenge and a wide area of research. For evaluations we used four descriptors of different types and sizes that have been produced by various partners of the IRIM project of the GDR ISIS [10].

- LIG_hg104: early fusion with normalization of an RGB histogram $4 \times 4 \times 4$ and a Gabor transformation (8 orientations and 5 scales), $64 + 40 = 104$ dimensions.
- CEALIST_global_tlep: early fusion of local descriptors of texture and of an RGB color histogram, $512 + 64 = 576$ dimensions.
- ETIS_global_qwm1x3x256: 3 histograms of 3 vertical bands of visual descriptors, standard Quaternion wavelet coefficients at three scales, $3 \times 256 = 768$ dimensions.
- LEAR_bow_sift_1000: histogram of local visual descriptors, SIFT “classic” [8], 1000 dimensions.

4.3 Optimal Negative to Positive Ratios

Table 2 shows the optimal values for the f_{pos} global parameter on the development set for single- and multiple-learner versions SVM-RBF with the four considered descriptors. Optimization was done while taking all the development set of TRECVID2007,

Table 2. Optimal values of the ratio between the numbers of negative to the positive samples for four descriptors

Descriptor	Single	Multi
LIG_hg104	4	2
CEALIST_global_tlep	8	4
ETIS_global_qwm	4	3
LEAR_bow_sift	8	4

in the Multiple-learner versions the results of the classifiers are fused by the harmonic mean function. These optimal values are higher for the single learner than for the multiple learner case. This was expected since the multiple-learner has another way to take into account more negative samples in total.

4.4 The Active Learning Steps

In our evaluation, we used totally 40 steps for the active learning algorithm, considering the geometric scale function with the following formula:

$$S_k = S_0 \times \left(\frac{N}{S_0} \right)^{k/K}$$

where N is the total size of the development set, S_0 is the size of the training set at the cold-start(30 samples), K is the total number of steps and k is the current step. At each step (or iteration) the algorithm calculates the S_k to be the size of the new training set and it chooses new samples to be labelled with size equal to $S_k - S_{k-1}$.

4.5 Active Learning Effectiveness

Figures 2 and 3 compare the effectiveness of the three methods (the single and Multi-learner and the Incremental) using the four descriptors and the relevance sampling strategy. The performance of the Single-learner is also shown as baseline. For the multiple-learner and the incremental experiments, the fusion by harmonic mean has been used. These plots show the evolution of the indexing performance of the test sets measured by the Mean Average Precision (MAP) metric with the number of annotated samples at each step (totally 40 steps). For analysing the plots we consider that The faster it grows and the higher performance it achieves , especially in the beginning, the better.

As the plots shown in figure 2 and 3, the proposed incremental algorithm has almost achieved the same performance as that of multi-learner. Both of them are significantly higher and faster to reach the highest value than single learner. With our incremental learning method, the highest performance can be reached with training by only 15-30% samples which means labelling 15-30% samples instead of all the samples can still get the same result (in MAP).

We considered the index of G_{a-b} to be the performance measure between two active learning curves(a and b), this measure was calculated simply as following: $G_{a-b} = (A_a - A_b)/A_b$ where A_i indicts the area of the curve i , to calculate this gain we

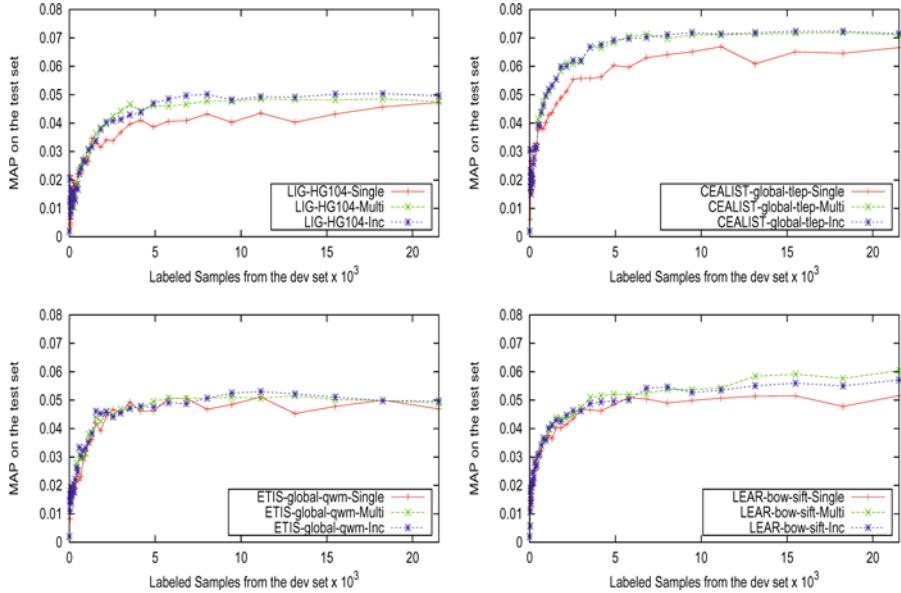


Fig. 2. The Map results on the TRECVID 2007 test collection evaluated on the four descriptors, each one of the plots shows the results using the Single-learner (in red), the Multi-learner (in green) and the Incremental (in blue)

first normalize the curves in each plot, then we calculate the area using the following formula:

$$A = \frac{1}{2} \left| \sum_{i=0}^{n+1} x_i \times y_{i+1} - y_i \times x_{i+1} \right|$$

Where n is the total number of iterations (x_i, y_i) indicates the (number of the annotated samples and the MAP value) at iteration i , and $(x_{n+1}, y_{n+1}) = (x_0, y_0)$. Table 3 show the gain when using the incremental method compared to both the single and multi-learner methods with the two collections that considered in our experiments, as we can see that the gain is much higher and significant when using our incremental method compared to the single-learner G_{I-S} while it is very small compared to the multi-learner G_{I-M} .

4.6 Execution Times

Table 4 gives the total execution times for the whole active learning process (40 iterations) on all 20 concepts on each experiment collection, per method and per descriptor, using the relevance strategy. As we can see that Single learner is faster than multiple learner and incremental methods. But considering the performance of single learner described in the above section, the performance of single learner is much lower than that of multi-learner. Compared with multiple learners, the new proposed incremental method has saved nearly 48-66% time without losing any performance.

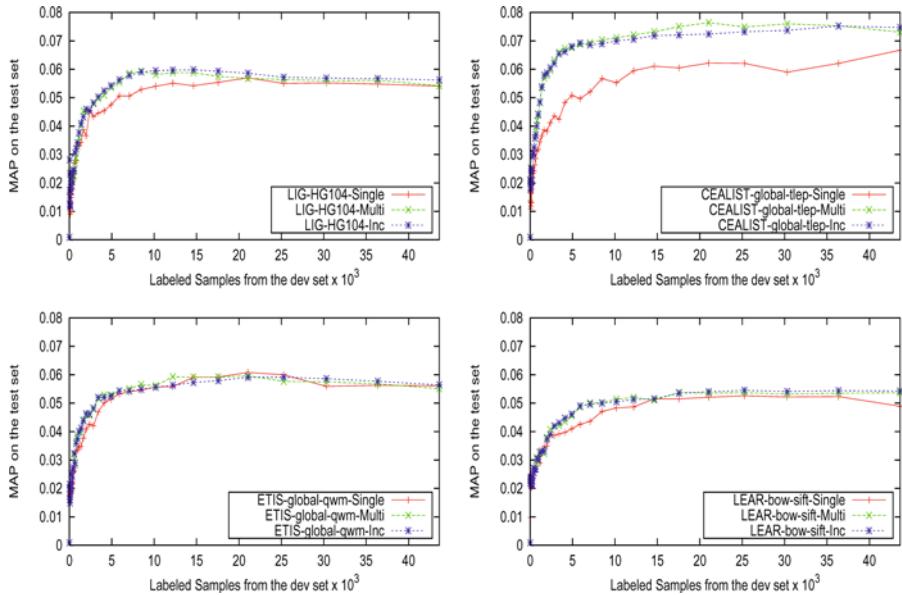


Fig. 3. The Map results on the TRECVID 2008 test collection evaluated on the four descriptors, each one of the plots shows the results using the Single-learner (in red), the Multi-learner (in green) and the Incremental (in blue)

Table 3. The gain of the system performance between the proposed incremental to the single- and the multi-learners, with the four descriptors evaluated on TRECVID 2007 and 2008

Descriptor	Trecvid 2007		Trecvid 2008	
	$G_{I-S}(\%)$	$G_{I-M}(\%)$	$G_{I-S}(\%)$	$G_{I-M}(\%)$
LIG_hg104	14.77	2.34	6.50	1.83
CEALIST_global_tlep	12.84	0.62	22.42	-1.70
ETIS_global_qwm	4.76	0.73	1.20	0.02
LEAR_bow_sift	8.04	-3.16	5.22	0.75

Table 4. Processing time table for the two evaluated collections: TRECVID 2007 and 2008, with G that indicates the gain of time using our incremental method compared to the multi-learners

Descriptor	Trecvid 2007				Trecvid 2008			
	Single	Multi	Inc	G	Single	Multi	Inc	G
LIG_hg104	1.40	20.63	7.64	66%	4.80	59.54	23.34	60%
CEALIST_global_tlep	23.90	115.02	64.17	52%	96.56	395.45	204.9	48%
ETIS_global_qwm	13.40	142.97	64.10	55%	45.67	460.60	212.3	54%
LEAR_bow_sift	43.42	162.18	79.16	52%	181.00	592.10	300.6	49%

5 Conclusion

Active learning with multiple classifiers has shown good performance for concept indexing in images or video shots in the case of highly imbalanced data. It involves however a large number of computations. In this paper, we propose a new incremental active learning algorithm based on multiple SVM for image and video annotation. The experimental result show that the best performance (MAP) is reached when 15-30% of the corpus is annotated and the new method can achieve almost the same precision while saving 50 to 63% of the computation time.

Acknowledgements

This work was partly realized as part of the Quaero Program funded by OSEO, French State agency for innovation.

References

1. Angluin, D.: Queries and concept learning. *Machine Learning* 2, 319–342 (1988)
2. Ayache, S., Quénot, G.: Evaluation of active learning strategies for video indexing. *Image Commun.* 22(7-8), 692–704 (2007)
3. Ayache, S., Queñot, G., Gensel, J.: Image and video indexing using networks of operators. *EURASIP Journal on Image and Video Processing* 2007(4), 1–13 (2007)
4. Bishop, C.M.: *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st edn. Springer, Heidelberg (2007)
5. Bordes, A., Ertekin, S., Weston, J., Bottou, L.: Fast kernel classifiers with online and active learning. *Journal Of Machine Learning Research* 6, 1579–1619 (2005)
6. Ertekin, S., Huang, J., Giles, C.L.: Active learning for class imbalance problem. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2007, pp. 823–824. ACM, New York (2007)
7. Ertekin, S.E., Huang, J., Bottou, L., Giles, C.L.: Learning on the border: Active learning in imbalanced data classification. In: Proc. ACM Conf. on Information and Knowledge Management, CIKM 2007 (2007)
8. Lowe, D.: Distinctive image features from scale-invariant key points. *International Journal of Computer Vision* 60(2), 91–110 (2004)
9. Naphade, M.R., Smith, J.R.: On the detection of semantic concepts at trecvid. In: Proceedings of the 12th Annual ACM International Conference on Multimedia, MULTIMEDIA 2004, pp. 660–667. ACM Press, New York (2004)
10. Quénot, G., Delezoide, B., le Borgne, H., Moellie, P.-A., Gorisse, D., Precioso, F., Wang, F., Merialdo, B., Gosselin, P., Granjon, L., Pellerin, D., Rombaut, M., Bredin, H., Koenig, L., Lachambre, H., Khoury, E.E., Mansencal, B., Benois-Pineau, J., Jégou, H., Ayache, S., Safadi, B., Fabrizio, J., Cord, M., Glotin, H., Zhao, Z., Dumont, E., Augereau, B.: Irim at trecvid 2009: High level feature extraction. In: TREC2009 notebook, November 16-17 (2009)
11. Ruping, S.: Incremental learning with support vector machines. In: Proceedings of the 2001 IEEE International Conference on Data Mining, ICDM 2001, Washington, DC, USA, pp. 641–642. IEEE Computer Society Press, Los Alamitos (2001)
12. Safadi, B., Quénot, G.: Active learning with multiple classifiers for multimedia indexing. In: CBMI, Grenoble, France (June 2010)

13. Safadi, B., Quénot, G.: Evaluations of multi-learners approaches for concepts indexing in video documents. In: RIAO, Paris, France (April 2010)
14. Shilton, A., Palaniswami, M., Member, S., Ralph, D., Tsoi, A.C., Member, S.: Incremental training of support vector machines. *IEEE Transactions on Neural Networks* 16, 114–131 (2005)
15. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, MIR 2006, pp. 321–330. ACM Press, New York (2006)
16. Snoek, C.G.M., Worring, M., Hauptmann, A.G.: Learning rich semantics from news video archives by style analysis. *ACM Trans. Multimedia Comput. Commun. Appl.* 2(2), 91–108 (2006)
17. Tahir, M.A., Kittler, J., Mikolajczyk, K., Yan, F.: A multiple expert approach to the class imbalance problem using inverse random under sampling. In: Benediktsson, J.A., Kittler, J., Roli, F. (eds.) MCS 2009. LNCS, vol. 5519, pp. 82–91. Springer, Heidelberg (2009)
18. Tahir, M.A., Kittler, J., Yan, F., Mikolajczyk, K.: Concept learning for image and video retrieval: The inverse random under sampling approach. In: 17th European Signal Processing Conference, Eusipco 2009, August 24–28 (2009)
19. Wen, Y.-M., Lu, B.-L.: Incremental learning of support vector machines by classifier combining. In: Proceedings of the 11th Pacific-Asia conference on Advances in knowledge discovery and data mining, PAKDD 2007. LNCS, vol. 4426, pp. 904–911. Springer, Heidelberg (2007)
20. Wu, C., Wang, X., Bai, D., Zhang, H.: Fast incremental learning algorithm of svm on kkt conditions. In: Proceedings of the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2009, Washington, DC, USA, pp. 551–554. IEEE Computer Society Press, Los Alamitos (2009)
21. Zhang, Y., Wang, X.: A fast support vector machine classification algorithm based on karush-kuhn-tucker conditions. In: Sakai, H., Chakraborty, M.K., Hassani, A.E., Ślęzak, D., Zhu, W. (eds.) RSFDGrC 2009. LNCS, vol. 5908, pp. 382–389. Springer, Heidelberg (2009)

A Semantic Higher-Level Visual Representation for Object Recognition

Ismail El Sayad, Jean Martinet, Thierry Urruty, and Chabane Dejraba

University of Lille 1, Telecome Lille 1
Lille, France

{Ismail.elayad,jean.martinet,thierry.urruty,chabane.dejraba}@lifl.fr

Abstract. Having effective methods to access the images with desired object is essential nowadays with the availability of huge amount of digital images. We propose a semantic higher-level visual representation which improves the traditional part-based bag-of words image representation, in two aspects. First, we propose a semantic model to generate a semantic visual words and phrases in order to bridge the semantic gab factor. Second, the approach strengthens the discrimination power of classical visual words by constructing an mid level descriptor, Semantic Visual Phrase, from frequently co-occurring Semantic Visual Words set in the same local context.

Keywords: Object Recognition, Bag of visual words, Visual phrases, Semantic analysis.

1 Introduction

As the holy grail of computer vision research is to tell a story from a single image or a sequence of images, object recognition has been studied for more than four decades [6]. Significant efforts have been paid to develop representation schemes and algorithms aiming at recognizing generic objects in images taken under different imaging conditions (e.g., viewpoint, illumination, and occlusion). Within a limited scope of distinct objects, such as handwritten digits, fingerprints, faces, and road signs, substantial success has been achieved. Object recognition is also related to content-based image retrieval and multimedia indexing as a number of generic objects can be recognized [7]. Indeed, the quality of the retrieving, indexing and classification depends on the quality of the internal representation for the content of the image.

Bag-of-visual-words [9] has drawn much attention between other approaches in the *part-base image representation*. Analogous to document representation in terms of words in text domain, the bag-of-visual-words approach models an image as an unordered bag of visual words. These visual words don't possess any semantics, as they are only quantized vectors of sampled local regions. In addition to the semantic factor, what really distinguishes textual word from

visual word is the discrimination power. Hence, in order to achieve better image representation, the low discrimination of visual words and the semantic factor must be tackled.

We generate visual element set comparable to the text words, Semantic Visual Words (SVWs) and Semantic Visual Phrases (SVPs) in this paper. SVWs are defined as the individual classical visual words that they are likely describing certain objects or scenes. Similar to the semantic meaningful phrases in documents, SVPs are defined as the semantic, distinctive and commonly co-occurring sets of SVWs occurring in the same local context in images. The frequent co-occurring set of semantic Visual Words Candidates SVWCs are found using *association rules* extracted with the *Apriori* algorithm [1] which form the semantic Visual Phrases Candidates (SVPCs). The SVWs and SVPs are selected from SVWCs and SVPCs candidates according to a semantic model (to be discussed later in this paper). Intuitively, once established, SVWs and SVPs will lead to compact and effective image representation.

The remainder of the article is structured as follows: Section 2, we describe the method for constructing the classical visual words from images. In Section 3 we propose a new semantic model based on different latent concepts (visual and high latent concepts). We generate the SVPCs in Section 4. In Section 5, we construct the SVWs and SVPs. We introduce a vote-based classifier for object recognition in Section 6. We report the experimental results in Section 7, and we give a conclusion to this article in Section 8.

2 Classical Visual Word Construction

In our approach, we use the SURF [2] low-level feature descriptor which is 64 dimensional vector that describes the distribution of pixel's intensities within a scale-dependent neighborhood of each interest point detected by the Fast-Hessian. In addition to the SURF descriptor, we used another descriptor (Edge context descriptor) introduced by El sayad et al.[4]. It describes the distribution of the edge points in the same Gaussian (by returning to the 5-dimensional color-spatialVisual words candidates are created by clustering the fused feature vectors in order to form a visual vocabulary. Quantization of the features into visual words is performed by using a vocabulary tree in order to support large vocabulary size. The vocabulary tree is computed by repeated k-means clusterings that hierarchically partition the feature space.

This hierarchical approach overcomes two major problems of the traditional direct k-means clustering in cases where k is large. Firstly clustering is more efficient during visual word learning and secondly the mapping of visual features to discrete words is way faster than using a plain list of visual words. Finally, we map each feature vector of an image to its closest visual word. Therefore we query the vocabulary tree for each extracted feature, and the best matching visual word index is returned.

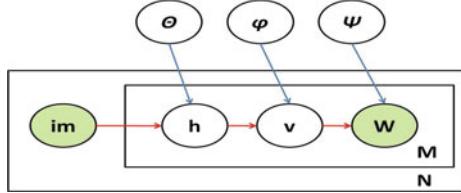


Fig. 1. The semantic model using the plate notation

3 Semantic Model for Generating the Semantic Visual Word Candidates (SVWCs)

3.1 Generative Process

Suppose that we have N Images $\{im_j\}_{j=1}^N$ in which M distinct classical visual word $\{w_i\}_{i=1}^M$ occur. We assume that every image consists of one or more visual aspects, which in turn are combined to higher-level aspects. This is very natural since images consist of multiple objects and belong to different categories. For this reason, we introduce two types of latent concepts (high latent concepts and visual latent concepts) in the following generative process for a given image im_j :

- Choose a high latent concept h_k from $P(h_k|im_j)$, a multinomial distribution conditioned on im_j & parameterized by $K \times N$ stochastic matrix θ , where $\theta_{kj} = P(h_k = k|im_j = j)$.
- Choose a visual latent concept v_l from $P(v_l|h_k)$, a multinomial distribution conditioned on h_k & parameterized by $L \times K$ stochastic matrix φ , where $\varphi_{lk} = P(v_l = l|h_k = k)$.
- generate a classical visual word w_i from $P(w_i|v_l)$, a multinomial distribution conditioned on v_l & parameterized by $M \times L$ stochastic matrix Ψ , where $\Psi_{il} = P(w_i = i|v_l = l)$.

$$P(w_i|im_j) = \sum_{k=1}^K \sum_{\ell=1}^L P(h_k|im_j, \theta) P(v_l|h_k, \varphi) P(w_i|v_l, \Psi) \quad (1)$$

$$L = \sum_{j=1}^N \sum_{i=1}^M n(w_i, im_j) \text{Log} P(w_i|im_j) \quad (2)$$

$$L = \sum_{j=1}^N \sum_{i=1}^M n(w_i, im_j) \text{Log} \left[\sum_{k=1}^K \sum_{\ell=1}^L P(h_k|im_j, \theta) P(v_l|h_k, \varphi) P(g_i|v_l, \Psi) \right] \quad (3)$$

3.2 Parameters Estimation

The expectation-maximization (EM) [3] maximum likelihood algorithm is the standard approach for maximum likelihood estimation in latent variable models.

The difficulty when implementing the EM algorithm is that a four dimensional matrix is required in the E-step because of two latent variables which costs high computational power. However, Gaussier et al. [5] proof that maximizing the likelihood can be seen as a nonnegative matrix factorization (NMF) problem under the generalized KL divergence. This leads to the following objective function:

$$\frac{\min}{\theta, \varphi, \Psi} GL(A, \Psi \varphi \theta) \quad (4)$$

where Ψ , φ , and θ are stationary points such that $\theta^T 1 = 1$, $\varphi^T 1 = 1$, $\Psi^T 1 = 1$

$$A_{ij} = n(w_i, d_j) \text{ and } GL(X, Y) = \sum_{i,j} (X_{i,j} \log \frac{X_{i,j}}{Y_{i,j}} - X_{i,j} + Y_{i,j})$$

This objective function can solved be by three-factor matrix factorization (also called the tri-NMF) where nonnegative constraints are imposed to all or only to the selected factor matrices: $\Psi \in \mathbb{R}_+^{M \times L}$, $\varphi \in \mathbb{R}_+^{L \times K}$, $\theta \in \mathbb{R}_+^{K \times N}$. If we do not impose any additional constraints to the factors (besides non negativity) which is the case here, the three-factor NMF can be reduced to the standard (two-factor) NMF [8] by the transformation $\varphi \theta \rightarrow \varepsilon$. However, the three-factor NMF is not equivalent to the standard NMF if we apply special constraints or conditions.

Firstly, we decompose to the Matrix A to Ψ and ε by using the multiplicative updating Rules to minimize (4). Lee and Seung [8] proposed a multiplicative algorithm to minimize (4) as following:

$$\varepsilon_{bj}^{k+1} = \varepsilon_{bj}^k \frac{((\Psi^k)^T(A))_{bj}}{((\Psi^k)^T(\Psi^k \varepsilon^k))_{bj}}, \forall b, j. \quad (5)$$

$$\Psi_{ia}^{k+1} = \Psi_{ia}^k \frac{(A(\varepsilon^{k+1})^T)_{ia}}{(\Psi^k \varepsilon^{k+1} (\varepsilon^{k+1})^T)_{ia}}, \forall a, i. \quad (6)$$

Such update rules do not always imply the convergence to a stationary point for these reasons:

- The denominator of the two Multiplicative updates rules may be zero.
- If ε_{bj}^k , numerator of the of one of the Multiplicative updates rules , is zero, and the gradient $\nabla_\varepsilon GL(A, \Psi \varepsilon) < 0$, ε_{bj}^{k+1} is not changed. Hence, it doesn't converges only at stationary point.

In order to overcome the two problems stated above, we propose modified multiplicative updating rules as the following:

$$\varepsilon_{bj}^{k,n} = \varepsilon_{bj}^k - \frac{\widehat{\varepsilon}_{bj}^k}{((\Psi^k)^T \Psi^k \widehat{\varepsilon}^k)_{bj} + \delta} \nabla_\varepsilon GL(A, \Psi^K \varepsilon^K)_{bj}, \forall b, j. \quad (7)$$

$$\Psi_{ia}^{k,n} = \Psi_{ia}^k - \frac{\widehat{\Psi}_{bj}^k}{(\widehat{\Psi}^k \varepsilon^{k,n} (\varepsilon^{k,n})^T)_{ia} + \delta} \nabla_\Psi GL(A, \Psi^K \varepsilon^{k,n})_{ia}, \forall a, i. \quad (8)$$

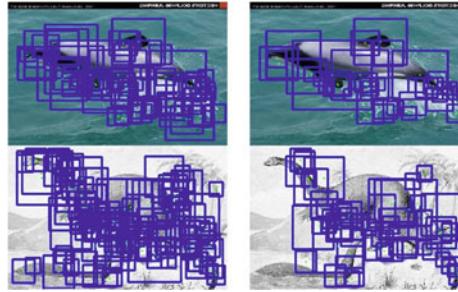


Fig. 2. The left side of the figure are images represented by the classical visual visual words and the right side of the figure are images represented by SVWCs

$$\text{where } \hat{\varepsilon}_{bj}^k \equiv \begin{cases} \varepsilon_{bj}^k & \text{if } \nabla_\varepsilon GL(A, \Psi^K \varepsilon^K)_{bj} \geq 0 \\ \max(\varepsilon_{bj}^k, \sigma) & \text{if } \nabla_\varepsilon GL(A, \Psi^K \varepsilon^K)_{bj} \leq 0 \end{cases} \quad (9)$$

Both δ and σ are pre-defined small positive numbers. Similarly, $\hat{\Psi}_{ai}^k$ is defined. Finally after estimating Ψ and ε . We estimate φ, θ by factorizing ε in the same procedure as above by applying the same multiplicative updates rules.

3.3 Semantic Visual Word Candidates (SVWCs) Generation

Before we generate the SVWCs, we categorize all the visual latent topics v_l according to their conditional probabilities with all the high latent topics $P(v_l/h_k)$. All the visual latent topics whose conditional probabilities to all the high latent topics are lower than a given threshold are categorized as irrelevant.

After that, all the classical visual words whose conditional probabilities $P(w_i|iv_l)$ are high to a given threshold for every relevant visual latent topic are categorized as semantic Visual Words Candidates (SVWCs). Figure 2 gives example of images represented by classical visual words and SVWCs within the same category.

4 Semantic Visual Phrase Candidates (SVPs)

Analogous to text documents, which are particular arrangements of words in 1D space, images are particular arrangements of patches in 2D space. The SVWs themselves and their inter-relationships generate candidates for the semantic visual phrases SVPs. We define SVP as a set of SVWCs that are frequently and coherently occurring together, with respect to certain semantic meaning.

Since it is not easy to define the semantic meaning of SVWC, we assume that set of SVWC are *semantically coherent* whenever they have high probability to the *same* latent visual latent topics. Their probability distributions are estimated using semantic model as defined in section 3.

4.1 Association Rules and SVPCs Generation

After describing the local context, we utilize the association rules to find the frequent item set with the SVWCs. Considering the set of all SVWCs (semantic visual words vocabulary) $W = \{w_1, w_2, \dots, w_k\}$ which denotes the set of items, D is a database (set of images I), $T = \{t_1, t_2, \dots, t_n\}$ is the set of all different sets of SVWs located in the same context which denotes the set of transactions.

An association rule is a relation of an expression $X \Rightarrow Y$, where X and Y are sets of items (sets of one or more of SVWCs that are within the same context). The quality of a rule can be described in the support-confidence framework. The support of a rule measures the statistical significance of a rule.

$$\text{support}(X \Rightarrow Y) = \text{supp}(X \cup Y) = \frac{|\{T_i \in T | (X \cup Y) \subset T_i\}|}{|T|} \quad (10)$$

The confidence of a rule measures the strength of the implication $X \Rightarrow Y$.

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(Y)} = \frac{|\{T_i \in T | (X \cup Y) \subset T_i\}|}{|\{T_i \in T | X \subset T_i\}|} \quad (11)$$

The left-hand side of a rule is called antecedent; the right hand side is the consequent. Note that the confidence can be seen as a maximum likelihood estimate of the conditional probability that X is true given that Y is true.

Given a set of images D , the problem of mining association rules is to discover all strong rules, which have a support and confidence greater than the pre-defined minimum support (*minsupport*) and minimum confidence (*min confidence*). Finally all frequent SVWC sets that occur in the same context and have the same semantic meaning are discovered which form the SVPCs.

5 Semantic Visual Word (SVW) and Semantic Visual Phrase(SVP) Generation

After generating the candidates for the SVWCs and SVPCs, both of them will form new observations for each image. We study the semantic meaning for the new observation in the same sense as we study the semantic meaning for the SVWC. We use the same semantic model that is introduced in section 3 but rather than using the visual words we as an observation, we use the semantic visual glossary candidate (a semantic visual glossary candidate can be a SVWC or SVPC) as the following:

- Choose a high latent topic h_k from $P(h_k | im_i)$, a multinomial distribution conditioned on im_j & parameterized by $K \times N$ stochastic matrix θ' , where $\theta'_{kj} = P(h_k = k | im_j = j)$.
- Choose a visual latent topic v_l from $P(v_l | h_k)$, a multinomial distribution conditioned on h_k & parameterized by $L \times K$ stochastic matrix φ' , where $\varphi'_{lk} = P(v_l = l | h_k = k)$.

- generate a visual glossary g_i from $P(g_i|v_l)$, a multinomial distribution conditioned on ω & parameterized by $M \times L$ stochastic matrix Ψ' , where $\Psi'_{il} = P(g_i = i|v_l = l)$.

$$P(g_i|im_j) = \sum_{k=1}^K \sum_{\ell=1}^L P(h_k|im_i, \theta') P(v_l|h_k, \varphi') P(g_i|v_l, \Psi') \quad (12)$$

$$L = \sum_{j=1}^N \sum_{i=1}^M n(g_i, im_j) \text{Log} P(g_i|im_j) \quad (13)$$

$$L = \sum_{j=1}^N \sum_{i=1}^M n(g_i, im_j) \text{Log} \left[\sum_{k=1}^K \sum_{\ell=1}^L P(h_k|im_i, \theta') P(v_l|h_k, \varphi') P(g_i|v_l, \Psi) \right] \quad (14)$$



Fig. 3. The images in the first row of the figure are images represented by the SVPC (They and are not selected later as SVP) while images in in the second row are images represented by SVPs

Maximizing the likelihood is done in the same manner as section 3. Finally, we generate the SVWs and SVWS from the semantic visual glossary candidates that have high probability to at least one relevant visual latent topic. The visual latent topics are categorized again after this new observation according to their probabilities to the high latent topics. 3 gives example of images represented by SVPCs and SVPs.

6 Vote-Based Classifier for Object Recognition

After we present a new image representation composed of semantic visual words (SVWs) and semantic visual phrases (SVPs) that allow us to consider the simple spatial relations between visual words, we propose a new vote-based classification technique relying on this representation. The object-based retrieving task is performed by utilizing this vote-based classifier. if most of the SVW and SVP candidates vote for high latent topic that is map to "Accordion" (the high latent topic and the class labels are mapped in the training set) then this image will

be recognized as "Accordion". In similar way, another two recognition results based on SVW only and SVP only are shown in the experiments section.

Firstly we estimate the voting score $VS_{h_k}^{SVW}$ for each high latent topic based on SVWs representation occurred in each test image im_j as the following:

- For each SVW_i occurred in an image im_j , we detect the relevant visual latent topic v that maximize the conditional probability $p(SVW_i|v_l)$.
- After detecting all the visual latent topics in image im_j , for each detected visual latent topic we detect the high latent topic h_k that maximize the conditional probability $p(v_l|h_k)$.

The final voting score $VS_{h_k}^{SVW}$ for each high latent topic based on SVW representation for an test image is

$$VS_{h_k}^{SVW} = \sum_{j=1}^{N_{im_j}} P(v_l|h_k) \quad (15)$$

where N_{im_j} is the number of high latent topics detected in im_j from the first step.

Secondly, In the same way, we estimate the voting score $VS_{h_k}^{SVP}$ for each high latent topic based on SVPs representation occurred on each in each test image im_j .

Finally, each image is categorized according to the dominant high latent topic which is the topic that has the highest voting score within the image as following:

$$VS_{h_k} = \alpha VS_{h_k}^{SVW} + (1 - \alpha) VS_{h_k}^{SVP} \quad (16)$$

α values are set according to contribute between the two representations.

7 Experiments

This section describes the set of experiments we have performed to explore the performance of the proposed methodology.

7.1 Dataset and Experimental Setup

The image dataset used for these experiments is the Caltech101 Dataset1. It contains 8707 images, which includes objects belonging to 101 classes. The number of images in each class varies from about 40 to about 800 with an average of 50 images. For the various experiments, we construct the test data set by selecting randomly 10 images from each class (1010 images). The visual word vocabulary size (K)=3000. and the query images are picked from this test data set during the experiment. For evaluation we estimate the performance for each object class as the following:

$$\text{class performance} = \frac{\sum_{i=1}^{N_{class}} N_{correct}}{N_{class}} \quad (17)$$

where N_{class} and $N_{correct}$ denotes the number of test images in a given class and number of test images that are correctly recognized respectively.

7.2 Contribution between the Classical Visual Words, SVWs and SVPs

In this section, we compare the performance of the classical visual word representations, SVW, SVP and SVW+SVP.

Firstly, features are extracted for each image in order to construct the classical visual words. classic visual word histogram is computed in each image, and histogram intersection is used as the distance metric for the bag of classical visual words approach. Each test image is recognized by computing its 10 nearest neighbors in the training dataset. Figure 4 plots the performance of the bag of classical visual words over the 101 object classes.

Secondly, we generate the SVWCs and SVPCS from the classical visual word then select the SVWs and SVPs in order to represent the images by two different representations (SVWs and SVPs). For each test image we run the vote-based classifier that we propose in section 6 by varying the parameter α introduced in the vote-based classifier. Figure 4 plots the object class performance for different values of α over all the 101 object. For a clear presentation, we arrange the 101 classes from left to right in the figure with respect to the ascending order of the SVWs+SVPs representation performance. When considering only SVPs ($\alpha = 0$), the performance is slightly better than the scenario in which only SVWs are used ($\alpha = 1$). It is obvious also that SVW representation is better than other scenario when using classical visual words over the 101 except in 5 categories. We notice that the average number of classical visual words in these categories is too small since the detected interest points were too small. Having few number of classical visual words leads to fewer number of SVWs that are selected from the classical visual words and this will affect the performance of the SVWs. However, the combination of SVWs and SVPs ($\alpha = 0.5$) yields better results than all other representations separately.

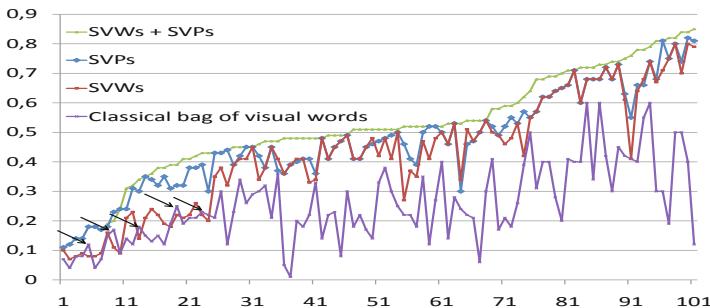


Fig. 4. Contribution Between the Classical Visual Words, SVWs and SVPs

7.3 Comparison between the Proposed Approach Performance and Similar Approaches

Following the experimental results in 7.2, we combine SVWs and SVPs voting for the object recognition task and we compare the performance of our

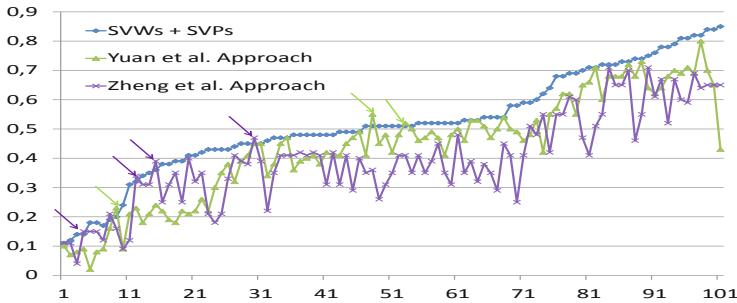


Fig. 5. Comparison between the proposed approach performance and similar approaches

proposed approach to two recent approaches based on classical visual words and phrases. Yuan et al. [10] proposed a higher-level lexicon, i.e. visual phrase lexicon, where a visual phrase is a meaningful spatially co-occurrent pattern of classical visual words. This higher-level lexicon is much less ambiguous than the lower-level one and likely describing an object or part of the object. [11] proposed a visual phrase-based approach to retrieve images containing desired objects. The visual phrase is defined as a pair of adjacent local image patches and is constructed using data mining. After constructing the two representations for these two approaches as proposed by Yuan et al and Zheng et al., respectively, the images are indexed with the visual words and phrases.

For these two approaches, TF-IDF weighting is applied to compute the similarities between images during the retrieval process. Each test image is recognized by computing the first 10 retrieved in the training dataset. Figure 5 plots the experimental results and for a clear presentation, we arrange the 101 classes from left to right in the figure with respect to the ascending order of the SVWs+SVPs representation performance. It is obvious that our proposed approach out performs the two other approaches in most of 101 classes([11] over performs our approach in 4 classes and [10] over performs on 3 classes).

8 Conclusion

In order to recognize objects within images, we proposed a semantic higher-level image representation. Firstly, we generated the semantic visual word candidates using a new proposed semantic model. Secondly, we exploited the spatial co-occurrence information of Semantic Visual Words candidates to generate a more distinctive visual configuration, i.e. Semantic Visual Phrase Candidates. Later, we select the Semantic Visual Words and Phrases that improve the discrimination power of the classical visual word representation and likely describe an object or part of object. Our experimental results have shown that the proposed approach over performs other recent approaches. In our future work, we will investigate the usage of such model on proposing computer vision solutions like

human behavior analysis from video. We will work on further justification based on other datasets.

References

1. Agrawal, R., Imielinski, T., Swami, A.N.: Mining association rules between sets of items in large databases. In: Buneman, P., Jajodia, S. (eds.) SIGMOD Conference, pp. 207–216. ACM Press, New York (1993)
2. Bay, H., Ess, A., Tuytelaars, T., Gool, L.J.V.: Speeded-up robust features (surf). Computer Vision and Image Understanding 110(3), 346–359 (2008)
3. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of The Royal Statistical Society, Series B 39(1), 1–38 (1977)
4. Elsayad, I., Martinet, J., Urruty, T., Djeraba, C.: A new spatial weighting scheme for bag-of-visual-words. In: International Workshop on Content-Based Multimedia Indexing (CBMI) (2010)
5. Gaussier, E., Goutte, C.: Relation between plsa and nmf and implications. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2005, pp. 601–602. ACM, New York (2005)
6. Lades, M., Vorbruggen, J.C., Buhmann, J., Lange, J., von der Malsburg, C., Wurtz, R.P., Konen, W.: Distortion invariant object recognition in the dynamic link architecture. IEEE Trans. Comput. 42(3), 300–311 (1993)
7. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2169–2178 (2006)
8. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: NIPS, pp. 556–562. MIT Press, Cambridge (2001)
9. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV, pp. 1470–1477. IEEE Computer Society, Los Alamitos (2003)
10. Yuan, J., Wu, Y., Yang, M.: Discovery of collocation patterns: from visual words to visual phrases. In: CVPR (2007)
11. Zheng, Q.-F., Gao, W.: Constructing visual phrases for effective and efficient object-based image retrieval. TOMCCAP 5(1) (2008)

Mining Travel Patterns from GPS-Tagged Photos

Yan-Tao Zheng¹, Yiqun Li¹, Zheng-Jun Zha², and Tat-Seng Chua²

¹ Institute for Infocomm Research, Singapore

² Department of Computer Science, National University of Singapore, Singapore
`{yzheng,yqli}@i2r.a-star.edu.sg, {zhazj,chuats}@comp.nus.edu.sg`

Abstract. The phenomenal advances of photo-sharing services, such as FlickrTM, have led to voluminous community-contributed photos with socially generated textual, temporal and geographical metadata on the Internet. The photos, together with their time- and geo-references, implicitly document the photographers' spatiotemporal movement paths. This study aims to leverage the wealth of these enriched online photos to analyze the people's travel pattern at the local level of a tour destination. First, from a noisy pool of GPS-tagged photos downloaded from Internet, we build a statistically reliable database of travel paths, and mine a list of regions of attraction (RoA). We then investigate the tourist traffic flow among different RoAs, by exploiting Markov chain model. Testings on four major cities demonstrate promising results of the proposed system.

Keywords: Geo-mining, human mobility analysis.

1 Introduction

The prevalence of photo capturing devices, together with the advent of media-sharing services like FlickrTM, have led to voluminous digital photos with text tags, timestamp and geographical references on the Internet. Different from other community-contributed multimedia data, these photos connect geography, time and visual information together and provide a unique data source to discover patterns and knowledge of our human society. In this study, our focus is to discover people's travel patterns within a local tour destination, by exploiting the geographically calibrated photos on photo-sharing websites, like Flickr¹. The rational is that the time-referenced and GPS-tagged photos implicitly document the spatio-temporal movements of their photographers. A large volume of such GPS-tagged photos can give rise to a statistical data source of people travel trails, as shown in Figure II.

Studies on people mobility and travel behavior within a local tour destination have always been important topics to mobile applications and location based services. In general, there exist two types of methods to acquire detailed travel

¹ <http://flickr.com>

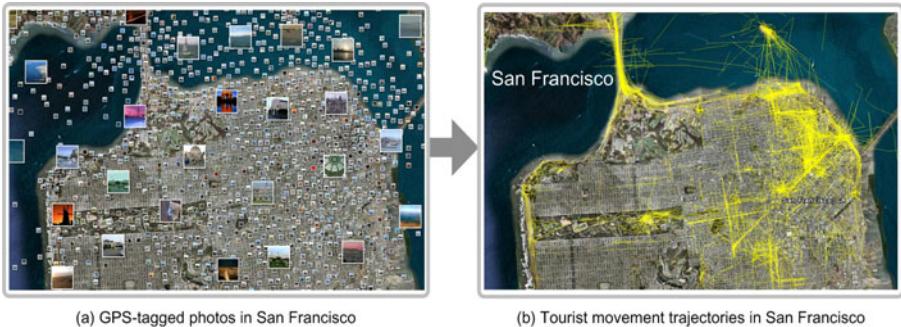


Fig. 1. GPS-tagged photos implicitly document the spatio-temporal movements of their photographers. The movement trajectories of photographers shown in (b) can be generated from GPS-tagged photos shown in (a). (For better viewing, please see the original color pdf file.)

data: (1) a survey with questionnaire on people's location histories [12]; and (2) location-acquisition devices for people to wear, such as GPS, cellular phone, etc, [19][12]. The issue with the first method is its expensive and time-consuming manual process, while the second method gives rise to unavoidable privacy issue that makes most people reluctant to participate in the study. In this study, we propose an Internet-driven approach to acquire people's travel information from GPS-tagged photos on the Internet. The advantage of this approach is that tourist mobility analysis can readily scale up onto a multitude of tour destinations. Such an automated travel pattern analytic approach can be tremendously useful to many geo-spatial applications. For example, the travel sequence analysis can reveal the crowd's choice of popular tour routes and help to monitor the traffic patterns of tourists.

To perform travel pattern analysis, we need first to build a statistically reliable database of people's travel paths. To do so, we discriminate tourist² photos v.s. non-tourist ones, based on the *mobility entropy* of a photo sequence pertaining to one photo uploader. The rational is simple: the mobile nature of sightseeing renders the photos of a true tourist to be spread over a large spatial extend within the tour destination. A Z-test is then applied on the movement trajectories generated from these tourist photos to ensure that the resulting photo trails are statistically reliable. Though outliers and noise might still exist, the travel patterns are expected to be statistically significant, when the number of trail samples is large. Figure 1 (b) depicts ~2000 mined photo trails in San Francisco. As shown, such large number of photo trail trajectories reveal conspicuous travel patterns. As our interest is to model the tourist mobility among different regions of attractions, we mine a list of *regions of attractions* (RoA) within a tour destination, by borrowing the approach in our previous work [18]. The premise is: a dense cluster of photos from different people indicate a region of frequent

² Tourist here is defined in a board sense to comprise both foreign and local people traveling for pleasure.

visits and popular appeal and is highly probable to be a region of attractions. After mining the list of RoAs, we represent tourist movement as a visit sequence of RoAs and exploit the Markov chain model [3] to analyze the tourist traffic statistics between different RoAs. The Markov chain model is widely used in various disciplines to analyze the trend of spatio-temporal movement and outcomes of sequential events [14]. Based on the first-order dependence in Markov chain, we estimate the statistics of visitors traveling from one region to another. Such tourist traffic analysis helps to indicate centric regions of attractions (RoA), which have influx of tourists from many other RoAs.

Overall, this study aims to exploit GPS-tagged photos on the Internet to analyze the travel patterns in a local destination. To the best of our knowledge, this is the first approach that leverages GPS-tagged photos for tourist traffic analysis. We demonstrate the proposed approach on four major cities in the world, i.e., Paris, London, San Francisco and New York City; and experiments show that the proposed approach can deliver promising results.

2 Related Work

In recent years, the advent of media-sharing services, such as FlickrTM and YoutubeTM, has led to voluminous community-contributed photos and videos available on the Internet. Together with socially generated textual and spatiotemporal metadata, these enriched multimedia data have spurred much research on discovering knowledge and patterns of our human society. Kennedy *et. al* proposed to discover aggregate knowledge of a geographical area, by analyzing spatiotemporal patterns of tags of Flickr photos in the area [8]. Similarly, Rattenbury *et. al* [13] and Yanai *et. al* [16] analyzed the spatiotemporal distribution of photo tags to reveal the inter-relation between word concepts (namely photo tags), geographical locations and events. Li *et. al* [10] and Zheng *et. al* [18] learned the geographical and visual appearance knowledge of tourist landmarks from community contributed photos on the Internet. The commonality between the aforementioned work and this study is that they all aim to extract some knowledge and patterns from photos with textual and spatiotemporal metadata, while the difference is that this study focuses on mining traveling patterns of tourists.

The study on tourist travel pattern within a tour destination has been a popular geographic research topic. McKercher and Lau [2] attempted to identify the movement patterns and styles of tourists within an urban destination. Asakura and Iryo [1] investigated the topological characteristics of tourist behavior in a clustering approach. Lewa and McKerchera [9] explored the urban transportation and tourist behavior modeling to identify explanatory factors that influence tourist movements. Compared to the work above, this study differs mainly in two aspects. First, the travel information in the previous work is mainly acquired via a manual survey with a limited number of tourist respondents. Consequently, the studies [2][1] covered only one or two tour destinations. In contrast, the proposed approach mines the travel information from Internet photos, which renders the data acquisition highly efficient, and thus, allows the travel analysis to easily

scale up to a multitude of destination. Second, constrating to existing approaches [12][1], this study analyzes the travel traffic by modeling it as sequence data via Markov chain model.

3 Approach

The overall framework consists of two major modules, i.e., building the travel path database and analyzing the travel traffic patterns.

3.1 Building the Travel Path Database

Given a set of GPS-tagged photos $\mathbb{P} = \{p\}$ within a tour destination, we build the database of travel paths. A photo p is a tuple $(\theta_p, \varphi_p, t_p, u_p, \varrho_p)$, containing the unique photo ID θ_p , tagged GPS coordinates φ_p in terms of latitude and longitude, time stamp t_p when photo was taken, photographer/uploader ID u_p and tagged text ϱ_p . Here, the tourist travel movement is modeled at a daily basis. According to photographer ID u_p , we organize photos of each photographer in a day in a chronological sequence $< p_0, \dots, p_k >$, which is defined as below.

Definition 1. *Photo sequence* of a photographer u_p is a chronological sequence $P = < p_0, \dots, p_k >$ of photos pertaining to the photographer u_p , where k is the number of photos and t_i is time stamp of photo i with $t_i < t_{i+1}$.

By representing the geographical calibration φ_p of photo p in ordinary Cartesian coordinates (x_p, y_p) , we define the movement of a photographer, in the notation of [5], as follows:

Definition 2. The *photo trail* of a traveler corresponds to a spatio-temporal sequence (ST-sequence) $S = < (x_0, y_0, t_0), \dots, (x_k, y_k, t_k) >$ drawn from chronologically ordered photo sequence P in a one-to-one corresponding manner, where $(x_i, y_i) \in \mathbf{R}^2$.

Based on Definitions 1 and 2, we construct the ST-sequence of movement trajectory of a photographer/uploader, by concatenating photos in the order of their time-stamp in a daily basis. We then classify these spatio-temporal sequences to tourist and non-tourist trails. The premise for classification is that the mobile nature of sightseeing renders the photos of a true tourist to be spread over a large spatial extend within the tour, as shown in Figure 2.

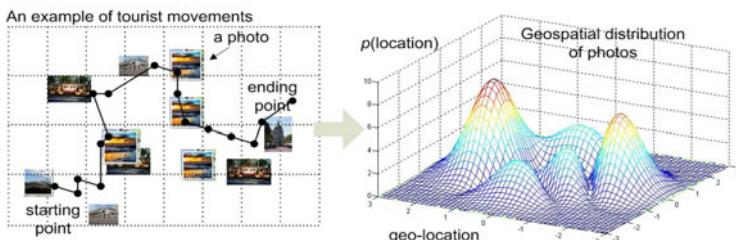


Fig. 2. Toy example of a tourist movement trajectory

Entropy based Mobility Measure. In a probabilistic perspective, the mobility complexity leads to a geospatial distribution of photos with reasonably high entropy. We, therefore, exploit this **mobility entropy** to discriminate the tourist and non-tourist movement trajectories, by utilizing the concept of Shannon entropy in Information Theory. Let $p(x, y)$ denote the geospatial density of photos with geospatial coordinates (x, y) pertaining to the photographer/upload. The mobility entropy $H_{mob}(S)$ of a movement trajectory $S = < (x_0, y_0, t_0), \dots, (x_k, y_k, t_k) >$ is computed as follows.

$$H_{mob}(S) = - \sum_i^n \sum_j^m p_{ij}(x, y) \log p_{ij}(x, y), \quad (1)$$

where $p_{ij}(x, y)$ is a discrete geospatial distribution of photos in grid (i, j) . By partitioning the tour destination into $n \times m$ grids, $p_{ij}(x, y)$ is estimated by the counts of photos in grid (i, j) . To discriminate photo trails, we empirically set an mobility entropy threshold ε_{mob} . The photo trail S is then classified as a tourist one, if $H_{mob}(S) \geq \varepsilon_{mob}$.

Statistical Significance of Travel Paths. We need to ensure that the resulting travel path database are statistically reliable. To do so, we perform statistical significance test on the resulting photo trails. Here, we characterize the photo trajectory of a photographer/upload with the number k of his/her visiting places. To some extent, the number of visit places indicates the mobility complexity of tourist. The number of visiting places is determined by the number of photos with unique geospatial coordinates. (Geospatial coordinates within a small distance will be considered as one.)

As the number of photo trajectory samples is relatively large, we approximate the tourist itinerary in terms of number of places k with normal distribution $\mathcal{N}(\mu_k, \sigma_k)$, based on *central limit theorem* in probability theory [2]. The mean μ_k and standard deviation σ_k are estimated from the trajectory samples. This normal model of tourist itinerary implicitly assumes individual tour itinerary samples are independent from each other. This independence assumption is reasonable, in the way that a tour itinerary depends on many factors, including tourist personal preference, his/her prior visits, total tour duration, tour destination demography, etc.

With the normal model, we apply Z-test to evaluate the statistical significance of a photo trajectory k . The null hypothesis H_0 in the test is defined as follows.

$$H_0 : k \sim \mathcal{N}(\mu_k, \sigma_k), \quad (2)$$

where H_0 states that the itinerary k generated from GPS-tagged photos is statistically significant, rather than generated from noisy photos by chance. The z-score z is then computed as below:

$$z = \frac{(k - \mu_k)}{\sigma_k / \sqrt{n}}, \quad (3)$$

where n is the number of photo itinerary samples. By looking up the z-score in a table of the standard normal distribution, the corresponding p-value can be

obtained. A lower p-value indicates a lower probability that the null hypothesis H_0 holds [17]. If p-value is less than a threshold τ , the null hypothesis H_0 is rejected and the photo itinerary is deemed to be statistically insignificant or unreliable and discarded subsequently.

Discovering Regions-of-Attractions (RoA). As our focus is on tourist mobility analysis at macro-level, a comprehensive list of regions of attractions within a tour destination is needed. Here, we define the region of attraction (RoA) as follows.

Definition 3. *A region of attraction r is a spatial extent in the geographical feature space of Cartesian coordinates (x, y) , where a considerable number of tourist movement trails pass through. RoA can be modeled as a spatial neighborhood function $F(x_i, y_i) : \mathbf{R}^2 \rightarrow \{0, 1\}$.*

In the spirits of our previous work [18], we develop a density-based model to discover regions of attractions, by analyzing the geospatial distribution of GPS-tagged photos. As stated in Definition 3, a region-of-attractions is a communal and interpretable spatial concept shared by a multitude of people. In other words, a RoA corresponds to a spatial extent, where many tourists visit and photograph. Clustering on GPS-tagged photos then become an intuitive solution to discover the list of regions of attraction.

Here, we adopt DBSCAN algorithm [4] to perform geospatial clustering on GPS-tagged photos for the following reasons. First, DBSCAN is a density-based clustering algorithm. Intuitively, it tends to identify regions of dense data points as clusters. This density driven approach just fits our task well, as the high density of photos implicates the popular appeal of the region. Second, DBSCAN algorithm supports clusters with arbitrary shape. This is critical to our task, as shapes of RoA can be spherical, linear, elongated etc. Third, DBSCAN is demonstrated to have good efficiency on large-scale data. (cf. [4] for more details of DBSCAN.)

After obtaining clusters of photos, we then determine the name and spatial extent of RoA, by examining the GPS coordinates and text title of component photos. We compute the frequency of n-grams of all photos text titles in each cluster. The name of RoA is determined to be the photo title with highest frequency. The geospatial extent of RoA is the area defined by the GPS coordinates of its member photos. Similar to [18], the resulting RoA is validated by the number of unique photographers/uploaders. This is to further ensure the popular appeal of RoA.

3.2 Transition Traffic between RoAs

Based on the concept of RoA, we define the *transition statistics between RoAs* as below.

Definition 4. *The transition statistics between RoAs depicts how tourist traffic flows from one RoA to another. It is defined as transition probabilities among different RoAs.*

By defining the tourist travel as a sequence of RoA, we investigate how tourists move from one RoA to another using the Markov chain model, in the spirit of [15].

In a statistical perspective, we model the movement of a tourist as an independent stochastic random process. The state space of the stochastic process is the set of RoA $\{r\}$ in the tour destination. Let $T = \{0, 1, 2, \dots\}$ denote the time index of the moves of a stochastic process. The stochastic process representing tourist movement $\{R_t\}_{t \in T}$ is referred to as a *Markov chain* (MC), if the value of next state does not depends on any previous states, given the value of current state, as defined below.

$$\begin{aligned} P(R_{t+1} = r_{t+1} | R_t = r_t, R_{t-1} = r_{t-1}, \dots, R_0 = r_0) \\ = P(R_{t+1} = r_{t+1} | R_t = r_t), \end{aligned} \quad (4)$$

where R_t is the random variable of RoA, r_t is a value of R_t and $r_t \in \{r\}$. In Markov chain model, each move in the state space $\{r\}$ is called a *step*. As each movement occurs after one unit time step, the stochastic process of tourist movement is modeled by a stationary discrete Markov chain. The transition probability $P(r_j | r_i)$ from RoA r_i to r_j can then be estimated by counting the tourists moving from RoA r_i to r_j . Accordingly, the RoA transition can be represented by a directed graph $G = (V, E)$, in which vertex V corresponds to RoA and edge E represents the transition statistics.

4 Experiments

GPS-tagged photos used in this study were downloaded from FlickrTM, by using its publicly available API [4]. To download photos, the name of a tour destination, such as Paris, London, etc, is fed in as query to retrieve a set of seed photos. Then, the owner ids of these seed photos are retrieved. Based on the owner ids, we download the entire collection of user's shared photos to ensure the completeness of the generated photo trail.

4.1 Travel Path Database

We download photos in four major cities: London, Paris, New York City and San Francisco. In total, we collected $\sim 769k$ GPS-tagged photos from $\sim 23k$ Flickr users. Based on Definition 1, we concatenate photos of a photographer into photo sequences in a daily basis. Following Section 3.1, we build a local travel database consisting of 8047 person-day trips by 5010 people in total. In average, each city has ~ 2000 person-day trips. This significantly outnumbers the manually collected tourist movement datasets of existing tourist mobility analysis works [12][11], not to mention that the database can be easily augmented by downloading more GPS-tagged photos. Figure 1 and 3 show the movement trajectories generated from GPS-tagged photos in New York City, San Francisco, Paris and London plotted on Google Earth.

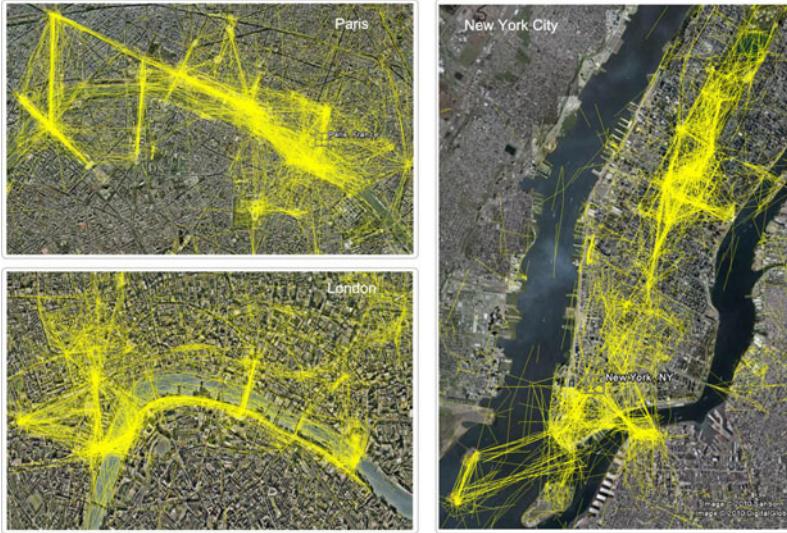


Fig. 3. Tourist travel trails generated from GPS-tagged photos in Paris and London

Regions of Attractions. By taking the photos in tourist movement trajectories as input, we discover the regions-of-attractions (RoAs) in a density-based approach, as presented in Section 3. In total, we discover 80 RoAs with 18 in London, 19 in Paris, 23 in New York City and 20 in San Francisco. Among them, only 1 out of 80 RoAs is false, which is "San Francisco Pride Parade". This event is misclassified as a RoA, as it gives rise to voluminous photos with strong geospatial pattern.

In the local travel database, the average cardinality of a daily trip routes is 3.5 RoA visits per day. This number is similar to the average visit of 3.7 RoAs per day in the tourism study [12].

4.2 Tourist Traffic Analysis

Popularity of RoA. The popularity of a RoA can be estimated by its tourist traffic volume, namely the number of people that have photographed in the region. Table II summarizes the top 3 most visited (most popular) RoAs in the four cities. For each RoA, the percentage of tourists that visit it is also computed. We compare Table II against the list of top 3 attractions in Yahoo!Travel [6] and found that two lists share 42% identical RoAs. The attraction popularity in Yahoo!Travel is estimated based on the feedback scores provided by Yahoo users. This overlap of popular RoAs suggests that despite of different background, people tend to agree on the most popular attractions to some extent.

Transition Traffic between RoAs. We utilize Markov chain model to estimate the transition probability $P(r_j|r_i)$. $P(r_j|r_i)$ indicates how tourist traffic

Table 1. Top three most visited RoAs and percentage of tourists in the four cities. SF: San Francisco.

	RoAs	Percentage (%)
SF	1. Golden Gate Bridge 2. Pier 39 3. Union Square	27.6 22.9 20.3
New York City	1. Times Square 2. Rockefeller Center. 3. Brooklyn Bridge	35.6 29.3 22.9
Paris	1. Notre Dame 2. Eiffel Tower 3. Arc de Triomphe	38.7 31.0 30.5
London	1. London Eye 2. Trafalgar Square 3. Tower Bridge	43.6 34.5 29.9

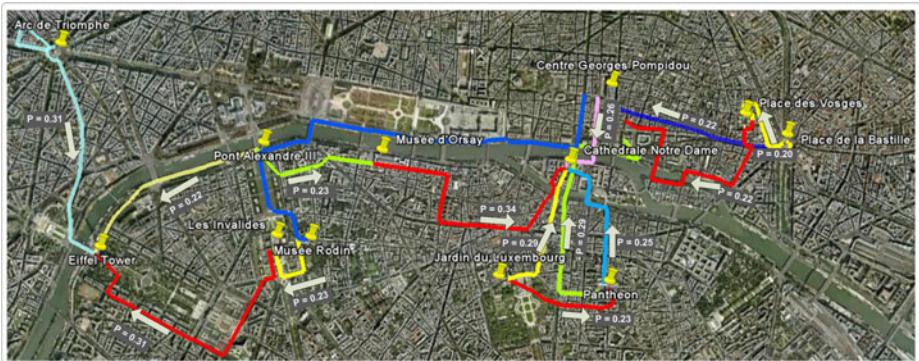


Fig. 4. Traffic transitions among RoAs in downtown, Paris, with transition probability $P(r_j|r_i) > 0.2$. For better viewing, please see the original color pdf.

moves from one RoA to another. A reasonably high value of $P(r_j|r_i)$ suggests that RoA r_j and r_i are coupled in the way that tourists tend to visit RoA r_j right after r_i . Figure 4 displays the RoA transitions with probability $P(r_j|r_i) > 0.2$ in downtown area of Paris. As shown, the coupled RoAs are usually geographically adjacent to each other. Moreover, it is also observed that people tend to prefer certain direction when visiting two coupled RoAs. For example, the transition probability $P(\text{Eiffel Tower} | \text{Arc de Triomphe})$ that tourists move from "Arc de Triomphe" to "Eiffel Tower" is 0.31, while the transition probability $P(\text{Arc de Triomphe} | \text{Eiffel Tower})$ in the opposite direction is only 0.12. This suggests that tourists might share similar preference in tour route planning. (For space limit reason, the RoA transition in the other three cities are not illustrated.)

Centric RoA. Tourist traffic tends to flow from several RoAs to a central one. We denote this central RoA as *centric RoA*. Specifically, we define centric

Table 2. Centric RoAs in the four cities

	Centric RoAs
San Francisco	Union Square, Chinatown
New York City	Time Square, Brooklyn Bridge
Paris	Eiffel Tower, Cathedrale Notre Dame
London	London Eye, Trafalgar Square

RoA as the one with transition probability $P(\text{centric RoA} \mid r_i) > 0.15$ for more than 3 RoA r_i . Table 2 summarizes the centric RoAs in the four cities. Figure 4 shows that "Eiffel Tower" and "Cathedrale Notre Dame" are centric RoAs in Paris, as they receive influx of tourists from several adjacent RoAs. The centric RoA might be determined by several factors, including popularity, geographical location, transportation convenience, etc. In a way, the centric RoA is the place where people congregate and meet each other.

5 Conclusion

Analysis on tourist mobility are important to tourism bureaucracy and industries. However, the cost of collecting detailed travel data is formidable. GPS-tagged photos available on the Internet implicitly provide spatio-temporal movement trajectories of their photographers. In this paper, we proposed to leverage these GPS-tagged photos to analyze the tourist travel behavior at the local level of a tour destination. We first built a statistically reliable tourist movement trajectory database from GPS-tagged photos, by utilizing an entropy-based mobility measure and Z-test. A list of regions of attraction (RoA) in a tour destination is then built, based on the frequency of tourist visits. We then investigated tourist traffic flow among different RoAs, by exploiting markov chain model to interpret tourists traffic transition. Finally, tourist travel patterns were analyzed by performing a sequence clustering on tour routes. Testing on four major cities, including San Francisco, New York City, Paris and London, demonstrated that the proposed approach can deliver promising results. One of our future works is to argument the tourist movement trajectory database and extend the travel analysis to a large scale of tour destination.

References

1. Asakura, Y., Iryo, T.: Analysis of tourist behaviour based on the tracking data collected using a mobile communication instrument. *Transportation Research Part A: Policy and Practice* 41(7), 684–690 (2007)
2. Degroot, M.H., Schervish, M.J.: *Probability and Statistics*, 3rd edn. Addison Wesley, Reading (2001)
3. Diaconis, P.: The Markov chain Monte Carlo revolution. *Bull. Am. Math. Soc., New Ser.* 46(2), 179–205 (2009)

4. Ester, M., Kriegel, H.-P., Jörg, S., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of Conference on Knowledge Discovery and Data Mining, USA, pp. 226–231. ACM, New York (1996)
5. Giannotti, F., Nanni, M., Pinelli, F., Pedreschi, D.: Trajectory pattern mining. In: Proceedings of Conference on Knowledge Discovery and Data mining, pp. 330–339. ACM, New York (2007)
6. Yahoo!Travel, <http://travel.yahoo.com/>
7. Flickr API, <http://www.flickr.com/services/api/>
8. Kennedy, L., Naaman, M., Ahern, S., Nair, R., Rattenbury, T.: How flickr helps us make sense of the world: context and content in community-contributed media collections. In: Proceedings of conference on Multimedia, pp. 631–640. ACM, New York (2007)
9. Lewa, A., McKercher, B.: odeling tourist movements: A local destination analysis. Annals of Tourism Research 33(2), 403–423 (2006)
10. Li, X., Wu, C., Zach, C., Lazebnik, S., Frahm, J.-M.: Modeling and recognition of landmark image collections using iconic scene graphs. In: Proceedngs of European Conference on Computer Vision, pp. 427–440 (2008)
11. McKercher, B., Lew, A.A.: Tourist flows and the spatial distribution of tourists. In: A Companion to Tourism, ch. 47, p. 36 (2004)
12. Mckercher, B., Lau, G.: Movement patterns of tourists within a destination. Tourism Geographies 10(3), 355–374 (2008)
13. Rattenbury, T., Good, N., Naaman, M.: Towards automatic extraction of event and place semantics from flickr tags. In: Proceedings of ACM SIGIR, pp. 103–110. ACM, New York (2007)
14. Upton, G.J.G., Fingleton, B.: Spatial Data Analysis Categorical and Directional Data, vol. 2. Wiley & Sons, Chichester (1989)
15. Xia, J.C., Zeephongsekul, P., Arrowsmith, C.: Modelling spatio-temporal movement of tourists using finite markov chains. Math. Comput. Simul. 79(5), 1544–1553 (2009)
16. Yanai, K., Kawakubo, H., Qiu, B.: A visual analysis of the relationship between word concepts and geographical locations. In: Proceeding of the ACM International Conference on Image and Video Retrieval, pp. 1–8. ACM, New York (2009)
17. Zha, Z.-J., Yang, L., Mei, T., Wang, M., Wang, Z.: Visual query suggestion. In: Proceedings of ACM International Conference on Multimedia, pp. 15–24 (2009)
18. Zheng, Y.-T., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T.-S., Neven, H.: Tour the world: building a web-scale landmark recognition engine. In: Proceedings of International Conference on Computer Vision and Pattern Recognition, Miami, Florida, U.S.A (June 2009)
19. Zheng, Y., Zhang, L., Xie, X., Ma, W.-Y.: Mining interesting locations and travel sequences from gps trajectories. In: Proceedings of the 18th International Conference on World Wide Web, WWW 2009, pp. 791–800. ACM, New York (2009)

Augmenting Image Processing with Social Tag Mining for Landmark Recognition

Amogh Mahapatra¹, Xin Wan¹, Yonghong Tian², and Jaideep Srivastava¹

¹ Department of CS, University of Minnesota, USA

² School of EE & CS, Peking University, China

{mahap022, wanxx061}@umn.edu

swantian@gmail.com

srivasta@umn.edu

Abstract. Social Multimedia computing is a new approach which combines the contextual information available in the social networks with available multimedia content to achieve greater accuracy in traditional multimedia problems like face and landmark recognition. Tian et al. [12] introduce this concept and suggest various fields where this approach yields significant benefits. In this paper, this approach has been applied to the landmark recognition problem. The dataset of flickr.com was used to select a set of images for a given landmark. Then image processing techniques were applied on the images and text mining techniques were applied on the accompanying social metadata to determine independent rankings. These rankings were combined using models similar to meta search engines to develop an improved integrated ranking system. Experiments have shown that the recombination approach gives better results than the separate analysis.

Keywords: Social Multimedia Computing, Landmark Recognition.

1 Introduction

A landmark can be defined as an identifiable location which has some kind of geographical, cultural or historical significance. There are many landmarks across the world which can be classified at various levels, e.g. city scale landmarks to international scale landmarks. In general, every important landmark can be identified with a representative set of images. For example, say a famous monument might have some popular front view images, side view images, night images etc.

Social networks are increasing the amount of multimedia content available online every minute. The number of pictures found on flickr.com for a search on the keyword Paris produces 174,047 results (on 8/15/2010), which is significant since this provides a tremendous image corpus which can be used as a starting point for recognizing Parisian landmarks. With increasing amount of multimedia content present online, the challenge now is to analyze, index and retrieve this content. The images are usually of varying quality, illumination, social relevance and context. The accompanying social metadata is often highly noisy and inaccurate, and the number of images also keeps increasing every day.

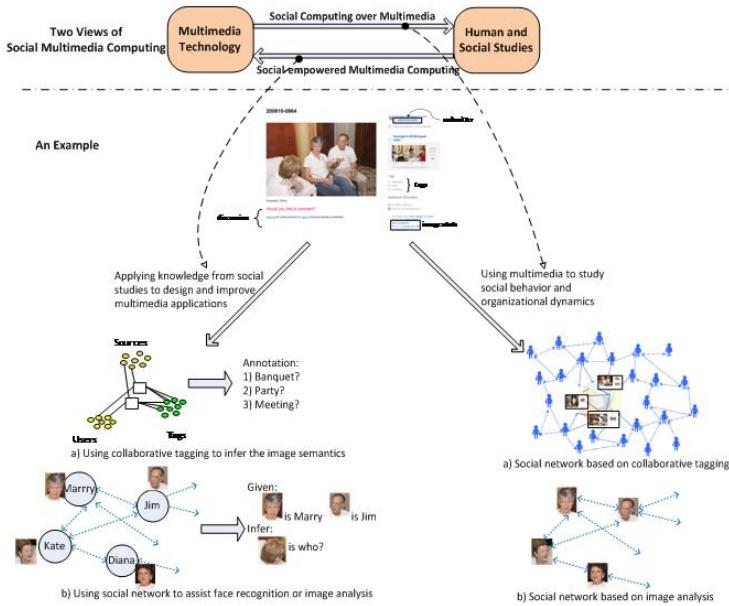


Fig. 1. The Social Multimedia Computing approach [12]

Hence online social networks, despite being rich sources of information, pose the difficult problem of information retrieval according to relevance. (Figure 1).

We will now review the work closely related to ours. Tsai et al. [1], found the context of a given set of images for a pre-defined landmark using GPS info and cell ids, and then did content analysis on images using SIFT [3] features. Their work established that context and content based analysis of social multimedia can improve the accuracy of image recognition. Simon et al. [10] use an unsupervised learning approach to create an image browser. They start with a set of images for a given location, cluster the images based on their visual properties using SIFT and create a browser where scenes can be explored. From each cluster they extract the most representative tags of the images. Their method puts forward a way of finding the most representative user tags. Naaman et al. [4] extract the landmark names across a given area using clustering based on geo-tags, from which they then extract the most representative tags for a given geo cluster. Next they extract the images corresponding to these tags and use image processing algorithms (like SIFT) on these images to find the best images. They used human experts for evaluation. Their work establishes that landmark names can be extracted using geo-tags and the use of a representative user tags as a preprocessing step can further improve the accuracy. Yan et al. [8] use a model where they first detect landmarks, and then filter images using the most representative tags. Finally they use image processing algorithms to mine and recognize landmark images. They created a web scale landmark recognition engine.

In this paper we propose a new approach which starts with a given set of images of a landmark and then clusters and ranks the images based on its visual features. Next an analysis of the tags and other social data is carried out separately. We next combine the rankings obtained from vision analysis and social data analysis using models similar to meta-search engines. In addition to proposing a new approach, we also include other available metadata in addition to geo-tags and user tags like Number of Views and Interestingness in our analysis. We believe the inherent bias in social data and the images can be properly handled using this approach.

2 Problem Statement and Proposed Approach

The Problem Statement: The problem is that of finding a set of images which can visually describe a landmark and which are diverse and yet precise at the same time. The accuracy of this image selection process is measured by how well these images can train a recognition system. If I is the set of images $I = \{i_1, i_2, \dots, i_n\}$ then the problem can be defined as that of finding a set of representative images I_s such that the following conditions hold, $I_s \in I, |I| >> |I_s|$.

The Proposed Approach: The images and the accompanying social data (e.g., geo-tags, user tags, number of views etc.) were analyzed separately after some initial pre-processing. The rankings generated by these different review techniques were combined using combination models similar to those used by meta search engines to obtain an integrated ranking. An image recognition system was designed to validate our tests. A test bed of images was selected, which contained a set of representative yet dissimilar images of a given landmark. The key point of our proposed approach is that instead of sequentially analyzing images and social data, we analyze them separately. We believe that this approach can handle the inherent bias present in the data extracted from the social networks. Rather than just using user tags and geo-tags we also explored other measures like number of views and Interestingness measure provided by flickr which further improved the accuracy of the proposed approach.(Figure 2)

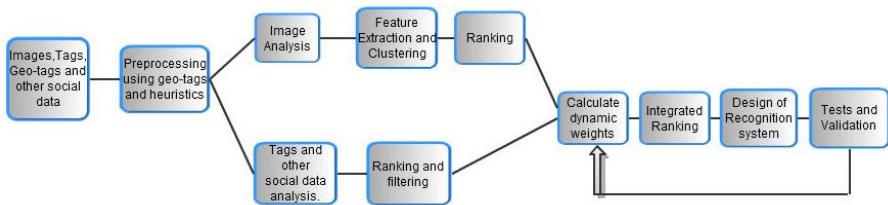


Fig. 2. The System Flow

3 Analysis of Images and Social Metadata

Our dataset is made up of the geo-tagged images downloaded from flickr.com. Some popular landmarks were selected randomly from across the world and their

corresponding images were downloaded from flickr.com (listed in Figure 5). For example, after selecting landmark X, flickr.com was queried using the keyword X. We downloaded 75,000 images for these 25 landmarks. Every image had the following metadata: Photo ID, Photo Owner ID, Title, Date taken, Date Upload, Tags, Geo-Tag(Latitude and Longitude), Views, Interestingness. Afterwards, the following heuristics were applied [4]:

1. The Photo-Owner ID was used to ensure that the number of unique uploaders for a given landmark was above a predefined threshold. This was to ensure that the number of photos included in our analysis did not end up belonging to just a few enthusiastic photographers.
2. The field date taken was used to ensure that the pictures used in our analysis did not belong to just one particular time frame; this was done to remove the local/cultural bias from the data. For example, a given landmark might have a lot of photos corresponding to a local festival.
3. Filtering using Geo Tags: Even after searching for a given keyword like Eiffel Tower from the flickr APIs the number of irrelevant images was still very high. The search results might include some popular restaurant, popular replicas etc. in other parts of the world. Hence, the latitude and longitude of these landmarks was used to filter out the irrelevant images.

3.1 Content Analysis of Images

The proposed image analysis has the following two steps. First, clustering of images was done based on the global features which results in a set of image clusters. This ensures that each cluster would contain different type of images e.g., one cluster could be of images of the landmark during the day while another could be of images taken at night. In the second step the images were ranked based on their local features. The inherent assumption is that the most representative images are similar to each other.

Clustering of Images: Same landmarks can have many images which could be very different from each other. Images could be taken from different parts of a building, from inside or outside, during day or night etc. It is hard to define the similarity between such images. At the same time, the representative image dataset should also include a diverse set of images, so that it can be used to train a recognition system which can recognize different images taken at different angles and illumination. Hence, the first step is to cluster the raw images. Two techniques were used for image clustering, Color Histogram [1] and Gabor features [2].

Color Histogram is a color quantization algorithm based on subjective vision perception. This model first transforms RGB values into HSV, and then describes the characteristics of an image using a histogram. It has relatively lower time and space complexity. For our experiments, H was divided into 8 parts, S and V were divided into three parts each. Hence, the histogram had 72 bins. Gabor wavelets are used to detect the texture features of the images. Mean and standard deviation based on six orientations and four frequencies were used to describe

the Gabor features, which ended up being a forty-eight dimensional array. The color and texture features together result in a 120 dimensional array for each image. Then, K-means algorithm was used to cluster these images into a certain number of groups. Thus, each cluster of images represents different views of the landmark.

Ranking of Images: More often than not the most representative images are visually similar. This implies that they should have a high degree of correlation. Thus the problem of finding the most representative images is the same as finding the images having high similarity with others in the same cluster. We used Scale Invariant Feature Transforms (SIFT) [3] to find the most representative images. SIFT is a local feature; it focuses on the objects in the images, and is invariant to changes in scale, rotation and illumination. We applied the definition of point-wise correspondences [4] to detect the overlap of images. So, two images that focus on similar objects will have a higher number of match points. After extracting all the interesting points from one particular image using SIFT algorithm, all these descriptors are indexed using a k-d tree.

The Best Bin First algorithm [5] was used to search for the match points. The degree of similarity between two images was quantified by the average number of point wise correspondence between one image and all the other images in the same cluster.

3.2 Analysis of User Tags

Most users tag their pictures after uploading them on a social network; if the image happens to be good, it gets tagged by the users too. It has been investigated by [9] that every landmark has a set of representative tags. Some previous approaches have used user tags as a preprocessing step, in their analysis. [8]

Our proposed approach assumes that just one tag is insufficient to perfectly identify a landmark but a landmark can be better identified by a set of tags S. For example, while searching for the most representative images of Eiffel Tower, the tag Eiffel Tower does not give the best results but if we include the images having the set of tags {‘paris’, ‘france’, ‘eiffeltower’, ‘eiffel’, ‘tower’, ‘europe’, ‘toureffel’} then the dataset happens to be more diverse and the recognition accuracy is higher too. The analogy behind this assumption has been derived from the usability aspect of search engines. For example, if we search for the keyword Java on any popular search engine then we get search results about, java the island, java the restaurant and many things other than the programming language. Hence, an end user always uses a particular set of appropriate keywords to refine his search and get to the most representative set of documents.

Find the Most Representative Tags: Noisy and misspelt tags were filtered using a standard thesaurus. Our assumption is that the most representative tags, which are invariant across seasons and local trends, should have a frequency higher than a certain threshold. Hence, a histogram analysis of the tag frequencies followed by a validation feedback loop was used to pick the most representative tags. (Figure [3])

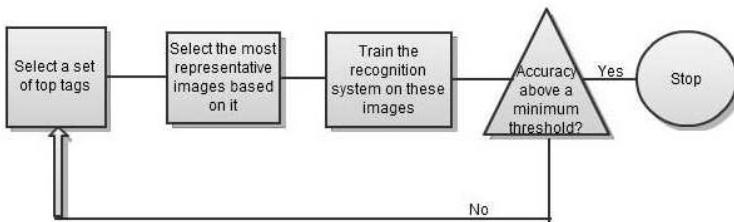


Fig. 3. Flowchart to select the best set of tags

Find the Most Representative Images: The TF-IDF model[13] was used to model the relationship between images and tags. Every image in our set was treated as a document and every tag was treated as word. Next, TF-IDF measure was used to normalize the matrix. The problem hence was reduced to finding the most relevant documents (images) that match our query vector which consists of the most representative tags as obtained previously. Latent semantic indexing was used to do this [14], where every image was treated as a vector in an n-dimensional space. Similarity of each vector was measured against the query vector and a final ranking was obtained.

3.3 Exploring the Number of Views

The data distribution of the number of views across images often gives a distribution where most images show almost the same number of views and some of them show abnormally high number of views(Figure 4). A curious observation is that the images with abnormally high number of views are usually not representative of the landmark; they usually consist of an artistic portrayal or a funny caricature like image. We found out that number of views is not a very good measure to find the most representative images. This was confirmed by training our recognition system on the most viewed images and the results turned out to be very poor.

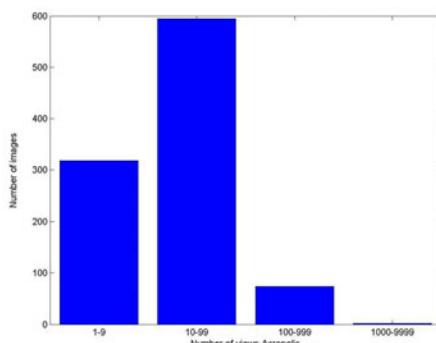


Fig. 4. No. of views of Images of Acropolis

3.4 Interestingness Measure

While downloading images from flickr.com one can specify the images to be sorted according to their interestingness. Interestingness [15] is a measure used by flickr.com to rate its images based on user comments, views, downloads etc. This measure was explored to see if it could yield representative images. After preprocessing and filtering the images we picked the most interesting ones from the remaining data set. Our recognition system was trained on these images and accuracy was measured. The results were ambiguous as the accuracy ranged from high accuracy to low accuracy. This is evident from the results given in Figure 5.

4 Combining Multiple Heterogenous Rankings

User tags, visual features and interestingness measure were used to rank the images independently. The rankings obtained by the above review techniques have different meaning, range and distribution. Hence, re-ranking models similar to those used by the meta-search engines were used to obtain an integrated ranking. The below models are not affected by the information and algorithm used by the review techniques.

The Agreement Model: The idea of the Agreement model [6] is that objects which have received a fair ranking in all reviews deserve a better ranking compared to the ones that have received a high rank in one and a low rank in another. The algorithm is as follows. Suppose we have m review techniques and n images.

1. W is the weighting vector.

$$W = [w_1, \dots, w_m]; \sum_{p=1}^m w_p = 1 \quad (1)$$

2. Agreement score of ith image (s_i) is computed as follows:

$$s_i = \sum_{j=0}^m \frac{w_j}{r_i^j} \quad (2)$$

r_i^j is the ranking of ith image in the jth review.

3. Sort images according to the score.

The Fuzzy Model: Fuzzy Model [7] uses the fact that some reviews might be more important than others in certain situations. Hence, it assigns unequal weights to different reviewers. The algorithm is as follows: The weights were decided both statically and dynamically, as described in the section 5.

1. Let S_i^j be the performance judgment of the ith image according to jth reviewer:

$$S_i^j = |L_j| - P_i^j + 1 \quad (3)$$

P_i^j is the position of ith image in the list L_j of jth reviewer.

2. Determine fitness score f_j which represents user preference of the jth reviewer $0 \leq f_j \leq L_j$
3. Define the weighting vector W(same as Equation 1) of IOWA operator [16]. The orness of the IOWA operator(for k reviewers) is defined by:

$$\text{orness}(W) = \frac{1}{k-1} \sum_{j=1}^k ((k-j) * (w_j)) \quad (4)$$

4. Calculate overall performance judgment S_i for each image by aggregating S_i^j using IOWA operator as follows:

$$S_i = \text{IOWA}(< u_i^1, S_i^1 >, \dots, < u_i^k, S_i^k >) \quad (5)$$

u_i^j is defined based on fitness score f_j of the jth reviewer and performance judgement S_i^j . Finally, sort images according to S_i^j .

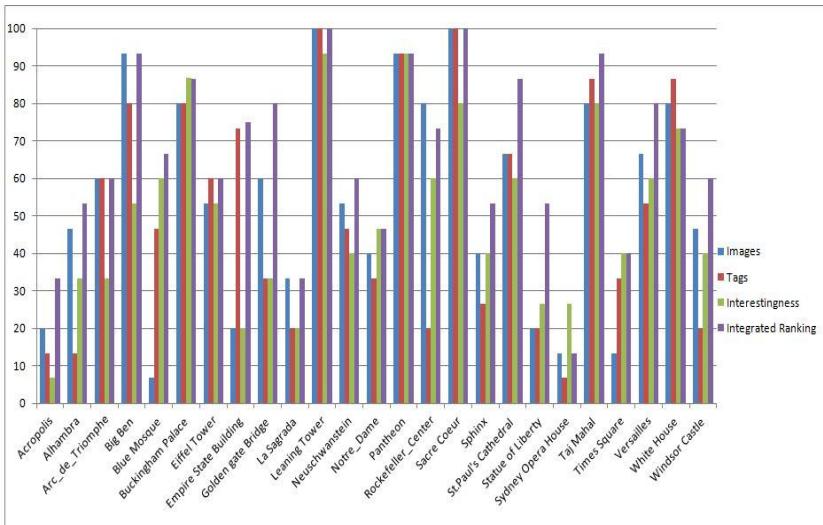


Fig. 5. Recognition accuracy for all landmarks, Y:Percentage, X:Landmarks

5 Evaluation of Proposed Approach

A recognition system was developed to validate the results. The SIFT features from each of the test images were extracted and compared with the k-d tree of each landmarks representative set of images. The landmark which had the highest number of match points with the test image was returned as the result. The system was trained using a variable number of images like the top 50,100, 200 images and it was found that a set of 75 images gave the best results. It was found that for certain landmarks the image analysis showed better results

compared to the social data analysis and vice-versa, hence a dynamic weighting system was adopted in our re ranking models where a particular review system was given a weight based on its observed recognition accuracy. If the recognition accuracy of first method of analysis was found to be A_1 and that of the second method was found to be A_2 then the weights were,

$$W_1 = \frac{A_1}{A_2 + A_1}; W_2 = \frac{A_2}{A_2 + A_1} \quad (6)$$

Six different models were used to integrate the ranks obtained from image and social data analysis as summarized below in the table.(Models 2 and 6 gave the best results respectively)

Models	Agreement Model	Fuzzy Model	Static Weights	Dynamic Weights	Images	Tags	Interestingness
Model 1	Yes	No	Yes	No	Yes	Yes	No
Model 2	No	Yes	Yes	No	Yes	Yes	No
Model 3	Yes	No	No	Yes	Yes	Yes	No
Model 4	No	Yes	No	Yes	Yes	Yes	No
Model 5	Yes	No	No	Yes	Yes	Yes	Yes
Model 6	No	Yes	No	Yes	Yes	Yes	Yes

For the experiments a set of 25 well known landmarks were selected.(listed in Figure 5). Afterwards a set of 75,000 images was downloaded from flickr.com using the landmark names as keywords.A set of diverse images not present in the training dataset was selected as the benchmark test set.

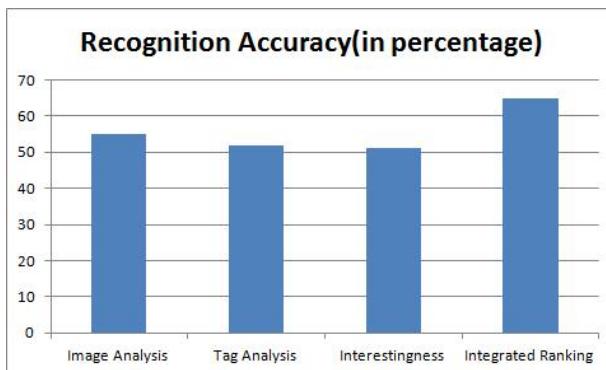


Fig. 6. Overall Recognition Accuracy,Y:Percentage,X:Lines of Analysis

It is evident from Figure 6 that the recognition accuracy increases after the process of integrated ranking. Some landmarks like Leaning Tower and Big Ben showed very high recognition accuracy because of their distinct visual features and popularity. The recognition accuracy was high as 95%. Some landmarks like Acropolis were a little more difficult to detect because of their spread out nature. Some landmarks have a similar skyline or periphery hence the system usually

confuses between them, like in this case, the system was usually confused between Statue of Liberty and Sydney Opera house, because of the accompanying sea. Hence, for better performance we might consider designing a recognition system which first asks for the users location then narrows down its search.

6 Conclusion

We have shown that landmark image mining and recognition can be improved by augmenting image processing with text mining techniques. A potential application of such system could be in various mobile devices where it could be used to recognize images taken by tourists. Such a system could also give tag recommendations to users. Future work, might investigate the origin and lifecycle of social data and incorporate them into solving traditional multimedia problems. Other useful social information other than tags could be used to gain better insights into the mechanics of community contributed web sites.

Acknowledgments

We would like to express our gratitude to members of the DMR lab in the University of Minnesota, for their valuable inputs during both design and discussion phase. This work is supported by a grant from the Chinese National Natural Science Foundation under contract number 60973055, a grant from the CADAL project and ARL Network Science CTA via BBN TECH/W911NF-09-2-0053.

References

1. Wan, H.L., Chowdhury, M.U.: Image Semantic Classification by Using SVM. *Journal of Software* 14, 1891–1899 (2003)
2. Zhang, D., Wong, A., Indrawan, M., Lu, G.: Content-based Image Retrieval Using Gabor Texture Feature. In: Proceedings of First IEEE Pacific-Rim Conference on Multimedia (PCM 2000), Sydney, Australia, pp. 392–395 (2000)
3. Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. *Int'l J. Computer Vision* 2(60), 91–110 (2004)
4. Kennedy, L.S., Naaman, M.: Generating diverse and representative image search results for landmarks. In: Proceeding of the 17th International Conference on World Wide Web, WWW 2008, pp. 297–306. ACM Press, New York (2008)
5. Beis, J., Lowe, D.G.: Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In: Conference on Computer Vision and Pattern Recognition, pp. 1000–1006, Puerto Rico (1997)
6. Oztekin, B., Karypis, G., Kumar, V.: Expert Agreement and Content Based Reranking in a Meta-search Environment Using Mearf. In: Proceedings of the 11th International World Wide Web Conference, pp. 333–344, Honolulu, Hawaii, USA, May 7–11 (2002)
7. Wiguna, W.S., Fernández-Tébar, J.J., García-Serrano, A.: Using a Fuzzy Model for Combining Search Results from Different Information Sources to Build a Metasearch Engine. In: International Conference 9th Fuzzy Days in Dortmund, Germany, pp. 325–334 (2006)

8. Zheng, Y., Zhao, M., Song, H., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T., Neven, H.: Tour the World: building a web-scale landmark recognition engine. In: Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR (2009)
9. Kennedy, L., Naaman, M., Ahern, S., Nair, R., Rattenbury, T.: How flickr helps us make sense of the world: context and content in community-contributed media collections. In: Proceedings of the 15th International Conference on Multimedia, Augsburg, Germany, September 25-29, pp. 631–640 (2007)
10. Simon, I., Snavely, N., Seitz, S.M.: Scene summarization for online image collections. In: Proceedings of the 11th IEEE International Conference on Computer Vision. IEEE, Los Alamitos (2007)
11. Tsai, C., Qamra, A., Chang, E.Y., Wang, Y.: Extent: Inferring Image Metadata from Context and Content. In: IEEE International Conference on Multimedia and Expo., pp. 1270–1273 (2005)
12. Tian, Y., Srivastava, J., Huang, T., Contractor, N.: Social Multimedia Computing. In: Computer IEEE Computer Society Digital Library, June 30. IEEE Computer Society, Los Alamitos (2010)
13. Ramos, J.: Using TF-IDF to Determine Word Relevance in Document Queries. In: First International Conference on. Machine Learning (2003)
14. Hoffman, T.: Probabilistic Latent Semantic Indexing. In: Uncertainty in Artificial Intelligence, UAI 1999, Stockholm (1999)
15. Explore About Interestingness, <http://www.flickr.com/explore/interesting>
16. Yager, R.R.: Induced aggregation operators. Fuzzy Sets and Systems 137, 59–69 (2003)

News Shot Cloud: Ranking TV News Shots by Cross TV-Channel Filtering for Efficient Browsing of Large-Scale News Video Archives

Norio Katayama, Hiroshi Mo, and Shin'ichi Satoh

National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-Ku, Tokyo 101-8430, Japan
{katayama, mo, satoh}@nii.ac.jp

Abstract. TV news programs are important target of multimedia content analysis since they are one of major information sources for ordinary daily lives. Since the computer storage cost has reduced significantly, today we can digitally archive a huge amount of TV news programs. On the other hand, as the archive size grows larger, the cost for browsing and utilizing video archives also increases significantly. To circumvent this problem, we present a visualization method of TV news shots using the popularity-based filtering across multiple TV channels. This method can be regarded as social filtering by TV broadcasters or popularity ranking among TV channels. In order to examine the effectiveness of our approach, we conducted an experiment against a thousand-hour order video archive storing 6 TV-channel streams for one month long. To our best knowledge, there is no former work applying this scheme to such a huge archive with conducting quantitative evaluation.

Keywords: News Shot Cloud, Video Archive, Cross TV-Channel Filtering.

1 Introduction

TV news programs are important target of multimedia content analysis since they are one of major information sources for ordinary daily lives. Since the computer storage cost has reduced significantly, today we can digitally archive a huge amount of TV news programs. On the other hand, as the archive size grows larger, the cost for browsing and utilizing video archives also increases significantly. To circumvent this problem, we present a visualization method of TV news shots using the popularity-based filtering across multiple TV channels. Our method is based on the assumption that if a news shot is informative, it would be more likely to be broadcasted by multiple TV stations. Obviously, news shots are major components of TV news programs but their roles are often limited to being supplementary to textual and audio information. On the other hand, as the proverb says, "seeing is believing". There exist the cases that news shots are much more informative than text and audio. If we want to fully utilize the power of TV news shots, we should focus on such informative cases. While informativeness of news shots might be evaluated by content analysis methods (i.e., news story parsing), we employ an alternative approach, i.e., an

accounts annex behavior bladecenter
 chiba company computer computing
 content data database dell dns
 dokewiki engine hardware
hitotsubashi homepage host hpc
 ibm index linux location machine nis
 node option pdf petastor plugins
 poweredge product remote research
 resource server share size smtp
software sun support syntax
system text unix wiki wysiwyg

Fig. 1. An Example of the Tag Cloud**Fig. 2.** An Example of the News Shot Cloud

information filtering approach, in which common news shots across multiple TV stations are automatically detected by video frame comparison. Then, the detected common shots are displayed in a diagram with changing the size of representative frames according to the number of occurrences across TV channels. We call this representation the "News Shot Cloud", since it can be regarded as a news shot version of the "Tag Cloud" that is widely used on the Web. Figure 1 shows an example of the tag cloud, while Figure 2 shows an example of the news shot cloud. The tag cloud provides a concise and intuitive sketch of the underlying document collection. Similarly, we can expect that the news shot cloud would be a concise and intuitive sketch of the underlying news shot collection. From another viewpoint, this method can be regarded as the social filtering by TV broadcasters or the popularity ranking among TV channels. To our best knowledge, there is no previous work applying the cross TV-channel filtering to the news shots ranking. In order to examine the effectiveness of our approach, we conducted an experiment against a thousand-hour order video archive storing 6 TV-channel streams for one month long. Our experiment demonstrates that the cross TV-channel filtering can be a strong method for browsing large-scale news video archives. To our best knowledge, there is no former work applying this scheme to such a huge archive with conducting quantitative evaluation.

This paper is organized as follows. In Section 2, related works are briefly presented. In Section 3, the cross TV-channel filtering method is presented, while the experimental results are shown in Section 4. Conclusions are in Section 5.

2 Related Works

In our proposed filtering method, common shots are detected by video frame comparison. The common shot detection has been vigorously investigated in recent years in the form of the copy or near-duplicate detection of video segments [4-12]. Our proposed method can be seen as one of its applications.

While this paper employs an information filtering approach for finding informative news shots, the standard approach is the content analysis approach, i.e., the news story parsing [1-3], in which news video is segmented into news stories based on visual and textual (speech) information. The news story parsing is essential to analyze the semantic details of news videos. On the other hand, the information filtering

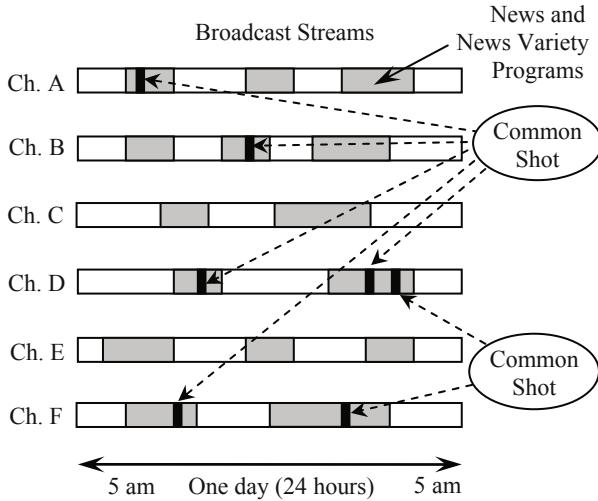


Fig. 3. Cross TV-Channel Filtering

approach is beneficial to obtain a rough sketch of news videos. We believe that both approaches are complementary to each other.

Similarly to our proposed method, some previous works conducts the common shot detection across multiple TV channels. Zhang et al. [4] detects near-duplicate images across news stories from two TV channels by part-based ARG (attributed relational graph) matching. Takimoto et al. [5] detects shots of the common scene by flash light patterns. Zheng et al. [7] presents an efficient near-duplicate keyframe detection method experimenting with the TRECVID-2006 corpus that covers 150hr news video from 6 TV channels with 3 different languages. Poullot et al. [9] presents a data mining framework based on the copy detection. Zhai et al. [3] detects common news stories across two TV channels and then ranks them according to the number of their occurrences within the two channels basing on the consideration that more "interesting" or "hot" story topics appear more times and longer than other stories. Thus, the common shot detection across multiple TV channels is not a new topic, but, to our best knowledge, no previous works have investigated the popularity ranking of news shots across multiple TV broadcasters.

3 Cross TV-Channel Filtering

3.1 Aim and Goal

Figure 3 illustrates the aim of the cross TV-channel filtering. By the advancement of the computer hardware capabilities, today, it is quite easy to construct a video archive storing a whole-day broadcast stream for multiple TV-channels. The aim of the cross TV-channel filtering is to find such news shots that are commonly broadcasted by multiple broadcasters. If we can assume that informative news shots are more likely

to be broadcasted by multiple broadcasters than others, we can expect that obtained common news shots are more informative than the others. In addition, we can rank them according to the number of their broadcasted TV channels. Figure 4 shows the flow of the cross TV-channel filtering. In the rest of this section, we describe the details of the flow.

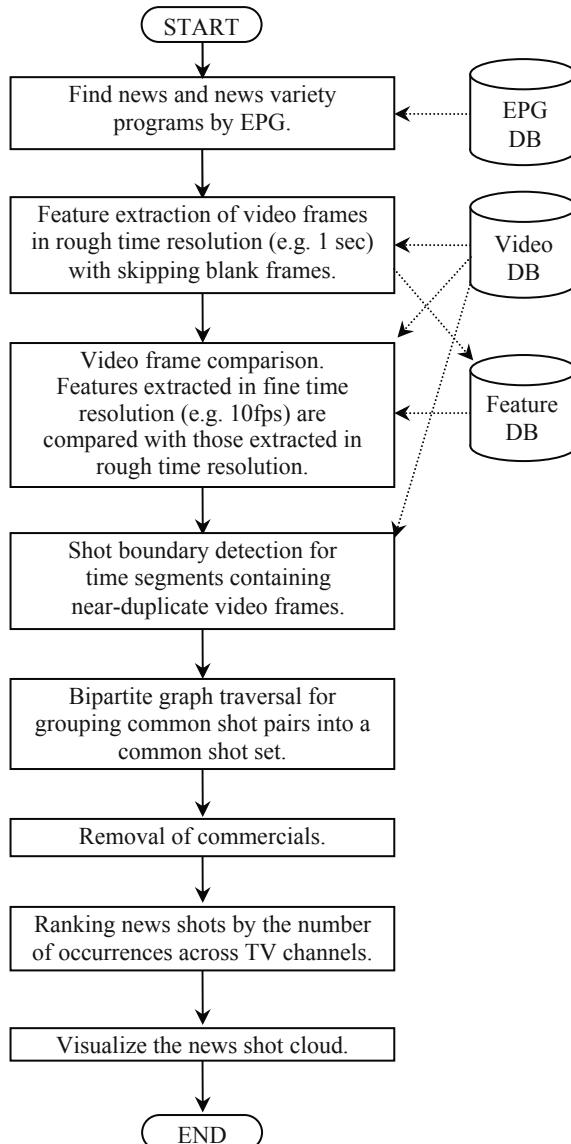


Fig. 4. Flow of Cross TV-Channel Filtering

3.2 Finding News Programs

Besides news programs, TV stations broadcast a wide variety of programs, e.g., dramas, sports, documentaries, etc. Therefore, in order to save the computational cost, we should restrict the processing targets to news and news variety programs. To that aim, we use the EPG (electrical program guide). The information contained in EPG depends on the EPG provider, but it is quite common that the category tag is assigned to each EPG. With such categorical information, we can select news and news variety programs from broadcast streams.

3.3 Video Frame Comparison

Once news and news variety programs are identified, the principle of vide frame comparison is very simple, i.e., compare every pair of video frames and then find near-duplicate pairs. However, in order to save computational cost, we employed the following strategies:

(1) Defining the minimum length of detecting shots.

This helps reducing the time resolution of video frame comparison. For example, if we define the minimum length of detecting shots to 1 second, we can compare two broadcast streams with picking every one-second frame for one broadcast stream. Please note that we should not pick every one-second frame for both two streams, because it may cause false dismissals if picking positions are shifted between two streams and if frame contents change dynamically within a shot.

(2) Detecting common shots that appear in at least two channels.

This means that we ignore near-duplicate shots that appear only in a single channel. This reduces the computational cost significantly. Without this strategy, common shots within a news program, e.g., anchor shots, diagrams, etc., are detected as near-duplicate shots. Please note that this does not exclude detecting such shots that appear repeatedly in a single channel. As long as they appear in at least two channels, such shots are also detected.

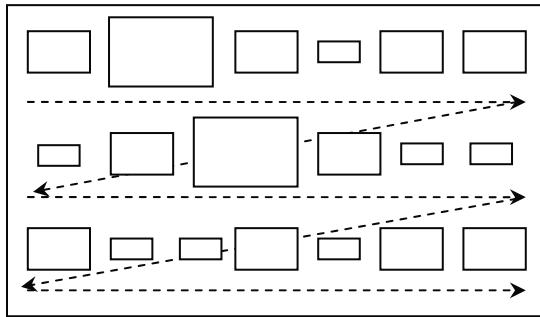
(3) Skipping blank frames.

Blank frames, i.e., blur, white-out, black-out, or dominated with some particular color, are ignored since they are less informative.

(4) Reducing the frame rate.

Since common shot detection is not sensitive to the time resolution, we reduced the frame rate from the standard 29.97fps to 9.99fps by picking every three frame for comparison.

Based on these strategies, video frame comparison consists of two stages: the former is the feature extraction in rough time resolution with skipping blank frames; and the latter is the comparison of two streams using pre-computed features for one stream and using features extracted in fine time resolution for the other.



Aligned in the order of the first appearance in the day.
For each shot, key frame is displayed with changing the size according to the number of broadcasted channels.

Fig. 5. Structure of News Shot Cloud

3.4 Shot Boundary Detection

Once near-duplicate frames are detected, shot boundary detection is applied to each time span that contains near-duplicate frames. This corresponds to assembling near-duplicate frame pairs into near-duplicate shot pairs. Shot boundaries are detected in a conventional way, i.e., the point where the frame content changes drastically is regarded as the shot boundary.

3.5 Grouping by Bipartite Graph Traversal

The near-duplicate shot pairs obtained by the comparison of two broadcast streams correspond to a bipartite graph in which two vertex sets represent the two broadcast streams. Each vertex represents a news shot in a broadcast stream, while each edge represents a detected shot pair. Since only such pairs that appear in different TV channels are detected, there exists no edge within a single broadcast stream. By finding the connected components in those bipartite graphs, we can group near-duplicate shot pairs into the sets of near-duplicate shots. The extraction of the connected components can be easily conducted by traversing bipartite graphs. Since the size of each component is not so large, the computational cost is not significant even with straightforward implementation.

3.6 Removal of Commercials

An annoying problem of the cross TV-channel filtering is that TV commercials are detected as the common shots. We employ some heuristics to distinguish news contents from TV commercials. The simple but effective heuristic is that the duration of TV commercials is usually standardized and much longer than that of common news shots. For example, in Japan, the duration of a TV commercial is a multiple of 15 seconds. On the other hand, it is rare that common news shots are broadcasted continuously for more than 15 seconds. Therefore, by finding a series of common shots that are broadcasted in a multiple of 15 seconds, we can assume that those shots are members of TV commercials.

Table 1. Ratio of News and News Variety Programs

Station	News	News Variety	Total
NHK	39%	11%	50%
NTV	27%	30%	57%
TBS	26%	23%	49%
FUJI	27%	18%	45%
ASAHI	32%	21%	53%
TOKYO	16%	25%	41%
Total	28%	21%	49%

Average from 08/31/2009 to 09/27/2009.

3.7 Visualizing News Shot Cloud

Finally, the obtained news shots are ranked by the number of occurrences across TV channels. Figure 5 shows the structure of the "News Shot Cloud". Similarly to the tag cloud, the size of keyframes are changed according to the number of broadcasters. There is no absolute rule for the alignment order of keyframes but this paper employs the order of the first appearance of the day because the time dimension is the most important domain for TV news programs.

4 Experimental Evaluation

In order to examine the effectiveness of our approach, we conducted an experiment against a thousand-hour order video archive storing 6 TV-channel streams for one month long. The cross TV-channel filtering method has been applied to each whole-day broadcast stream (6ch \times 24hr).

4.1 TV Broadcast Archive

In our experiment, we used a TV broadcast archive which contains broadcast streams of 6 terrestrial TV channels available in Tokyo, Japan. The recording format is MPEG1; the frame size is 352×240 , the frame rate is 29.97fps, and the bit rate is 1.5Mbps. Each channel is provided by an independent broadcast station. They are not a specialty channel and cover general topics. In each channel, TV programs are broadcasted almost 24 hours a day. One of 6 broadcast stations is a public broadcaster, NHK, and the others are private companies. They belong to different major TV networks in Japan and compete with each other to have more viewers. Thus, all 6 channels are suitable for cross TV-channel filtering. The number of available TV channels varies from city to city. Since six stations are quite many compared with the other cities, Tokyo is a good location for the experiment location. Table 1 shows how long each TV station broadcasts news or news variety programs. The total ratio of each station ranges from 41% to 57%; thus, each station broadcasts news and news variety programs almost a half of the day. Therefore, the total length of these programs reaches over 2,000 hours. It is quite a big challenge to apply cross TV-channel filtering to this archive.

4.2 Frame Comparison Method

Our current implementation of the frame comparison method is quite simple. Mainly due to its simplicity and low computational cost, we employ the conventional feature matching method. Our aim is to detect the same or very similar shots with allowing minor differences like the insertion of open caption texts around the borders, the difference of the camera viewpoints, the difference of color illumination, etc. To this aim, we employed the following feature vectors:

- DCT lower components of the pixel intensity,
- DCT lower components of the pixel color (along red axis in the HSI space),
- DCT lower components of the pixel color (along yellow axis in the HSI space),
- DCT lower components of the color gradient (magnitude and orientation).

The dimensionality of the DCT components was 36 (6×6).

Since the border area of each frame is vulnerable to the insertion of open caption texts, we computed the above feature vectors for the two areas of each frame; one is the whole frame area and the other is the center area of the frame. Then, we used a tighter threshold for the center area, while a looser threshold is used for the whole area. A pair of frames is regarded as almost the same if all of the above feature vectors are close to each other within the given threshold. Although the selectivity of each feature vector is not so high, their combination enhances the selectivity. As far as we experimented, the precision was satisfactorily high (error was less than 1%).

4.3 Computational Cost

Here, we present the rule of thumb of computational cost. The major cost is the video frame comparison. As mentioned above, the video frame comparison consists of two stages; one is feature extraction with rough time resolution and the other is feature vector comparison in fine time resolution. In our current implementation, the speed of the former stage is about 5fps, while the latter stage is about 4fps, using a single processor core of the ordinary Intel Xeon processor. In both stages, the dominant cost is the decoding of MPEG1 video stream. Thus, the total processing time of each stage for processing a whole day broadcast stream is as follows:

- Computing feature vectors in rough time resolution:

$$250,000 \text{ frames} (6\text{ch} \times 12\text{hr} \times 1\text{fps}) / 5\text{fps} = 14\text{hr}$$
- Feature vector comparison in fine time resolution:

$$2,500,000 \text{ frames} (6\text{ch} \times 12\text{hr} \times 10\text{fps}) / 4\text{fps} = 170\text{hr}$$

Table 2. Size of Matched Shots (Hour:Min)

Station	Shots Matched with Another Channel		
	News Shots	Commercial	Total
NHK	0:16	0:00	0:16
NTV	0:37	1:24	2:00
TBS	0:24	1:03	1:27
FUJI	0:29	1:05	1:33
ASAHI	0:36	1:13	1:48
TOKYO	0:13	0:58	1:11
Total	2:30	5:42	8:42

Result of 09/01/2009

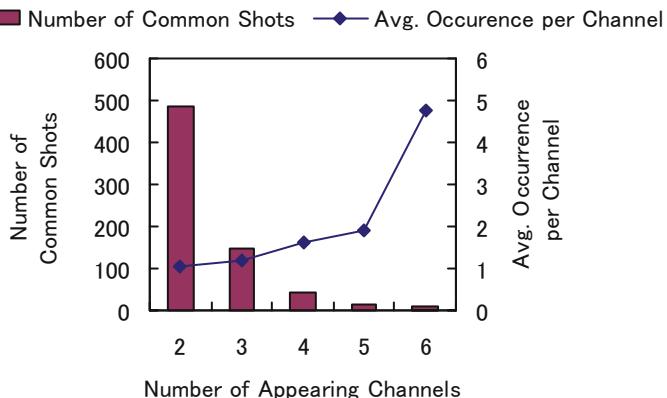


Fig. 6. Number of Common Shots vs Number of Appearing Channels (09/01/2009)

For 30 day broadcast stream, the total processing time is summed up to 5,500 hours. This figure clearly demonstrates that the cross TV-channel filtering is a big challenge. In our experiment, we used a PC cluster with 400 processor cores. Thus, we could complete the filtering just in a day.

4.4 Size and Quality of Selected News Shots

Table 2 shows the size of matched shots. From 144 hour broadcast stream ($6\text{ch} \times 24\text{hr} \times 1\text{day}$), the total duration of matched shots is about 8 hours. Within those shots, news shots are 2 hour, while commercial shots are 6 hour. Thus, the selectivity is about $2/144 \approx 1\%$. The size reduction ratio is quite high. On the other hand, Figure 6 shows the relationship between the number of common shots and the number of appearing channels. Naturally, the number of common shots decreases as the number of appearing channels increases. Meanwhile, the average occurrence per channel increases as the number of appearing channels increases. This means that the shots appearing in more TV-channels are more likely to be used in TV news programs. This observation is consistent with our assumption that if a news shot is informative, it would be more likely to be broadcasted by multiple TV stations.

For the cross validation, we employed a weekly news program broadcasted every Saturday morning by NHK which presents the major news topics of the week. Figure 7 shows the number of the common shots appearing both in the weekly news program and in the daily news programs of that week. By comparing Figure 6 and 7, it can be seen that the ratio of the common shots appearing in 6 channels much differs between the weekly and the daily programs. That of the weekly program is much greater than that of the daily programs. To clarify this observation, Figure 8 shows the relationship between the number of appearing channels and the difference in the percentages between the weekly and the daily news programs. As shown in Figure 8, the common shots appearing in more channels are more likely to be used in the weekly news program. This result also supports our assumption that if a news shot is informative, it would be more likely to be broadcasted by multiple TV stations.

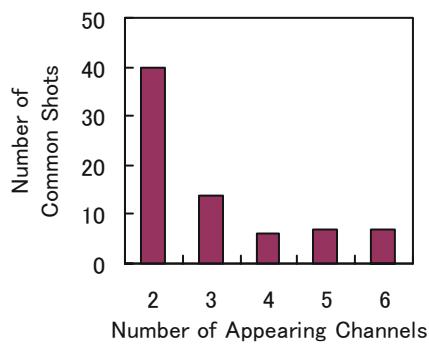
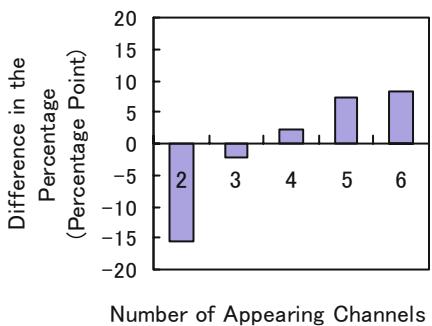
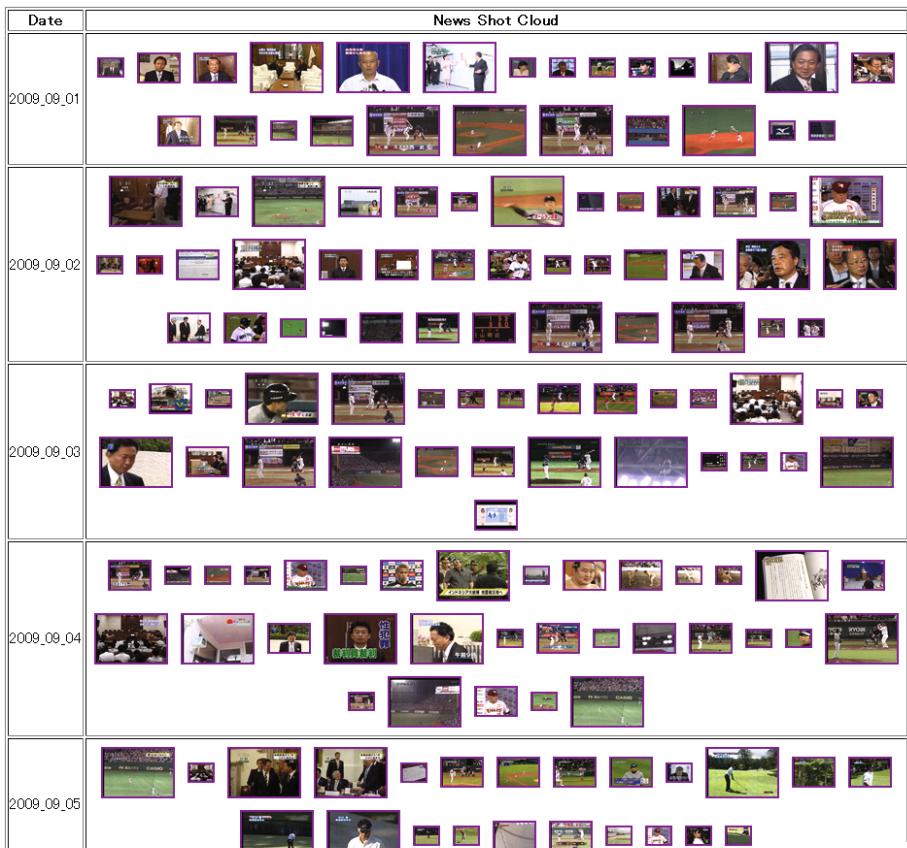
**Fig. 7.** Case of a Weekly News Program**Fig. 8.** Weekly vs Daily Compared in the Percentage of Common Shots**Fig. 9.** Calendar View of News Shot Cloud

Figure 9 shows a calendar view of the news shot cloud. In this view, the common shots appearing in 4 and more channels are displayed. This illustrates that the news shot cloud can be a concise and intuitive sketch of the underlying news shot collection.

5 Conclusions

In this paper, we presented a method for constructing a news shot version of the "Tag Cloud" called the "News Shot Cloud" based on the cross TV-channel filtering. This method can be regarded as the social filtering by TV broadcasters or the popularity ranking among TV channels. We conducted an experiment against a thousand-hour order video archive storing 6 TV-channel streams for one month long. Our experiment demonstrates that the cross TV-channel filtering can be a strong method for browsing large-scale news video archives.

Future works include as follows:

- Enhancement of computational efficiency.
- Usability study for evaluating the effectiveness of the news shot cloud.

Acknowledgments. This work was partially supported by JSPS KAKENHI (Grant-in-Aid for Scientific Research) No. 21300039, 19500102, and 20300041.

References

1. Haupmann, A.G., Witbrock, M.J.: Story Segmentation and Detection of Commercials in Broadcast News Video. In: Proceedings of the Advances in Digital Libraries Conference, pp. 168–179. IEEE Computer Society, Washington (1998)
2. Wu, C.-H., Hsieh, C.-H.: Story Segmentation and Topic Classification of Broadcast News via a Topic-Based Segmental Model and a Genetic Algorithm. *IEEE Trans. on Audio, Speech, and Language Processing*, 17(8), 1612–1623 (2009)
3. Zhai, Y., Shah, M.: Tracking News Stories across Different Sources. In: Proceedings of ACM Multimedia 2005, pp. 2–10. ACM, New York (2005)
4. Zhang, D.-Q., Chang, S.-F.: Detecting Image Near-Duplicate by Stochastic Attributed Relational Graph Matching with Learning. In: Proceedings of ACM Multimedia 2004, pp. 877–884. ACM, New York (2004)
5. Takimoto, M., Satoh, S., Sakauchi, M.: Identification and Detection of the Same Scene based on Flash Light Patterns. In: Proceedings of IEEE ICME 2006, pp. 9–12. Pergamon Press, Inc., Elmsford (2006)
6. Wu, X., Hauptmann, A.G., Ngo, C.-W.: Novelty Detection for Cross-Lingual News Stories with Visual Duplicates and Speech Transcripts. In: Proceedings of ACM Multimedia 2007, pp. 168–177. ACM, New York (2007)
7. Zheng, Y.-T., Neo, S.-Y., Chua, T.-S., Tian, Q.: The Use of Temporal, Semantic and Visual Partitioning Model for Efficient Near-Duplicate Keyframe Detection in Large Scale News Corpus. In: Proceedings of the ACM International Conference on Image and Video Retrieval 2007 (CIVR 2007), pp. 409–416. ACM, New York (2007)
8. Wu, X., Takimoto, M., Satoh, S., Adachi, J.: Scene Duplicate Detection Based on the Pattern of Discontinuities in Feature Point Trajectories. In: Proceeding of ACM Multimedia 2008, pp. 51–60. ACM, New York (2008)

9. Poullot, S., Crucianu, M., Buisson, O.: Scalable Mining of Large Video Databases Using Copy Detection. In: Proceedings of ACM Multimedia 2008, pp. 61–70. ACM, New York (2008)
10. Tan, H.-K., Ngo, C.-W., Hong, R., Chua, T.-S.: Scalable Detection of Partial Near-Duplicate Videos by Visual-Temporal Consistency. In: Proceedings of ACM Multimedia 2009, pp. 145–154. ACM, New York (2009)
11. Zhou, X., Zhou, X., Chen, L., Bouguettaya, A., Xiao, N., Taylor, J.A.: An Efficient Near-Duplicate Video Shot Detection Method Using Shot-Based Interest Points. IEEE Trans. on Multimedia 11(5), 879–891 (2009)
12. Döhring, I., Lienhart, R.: Mining TV Broadcasts for Recurring Video Sequences. In: Proceedings of the ACM International Conference on Image and Video Retrieval 2009 (CIVR 2009), no. 28. ACM, New York (2009)

Speaker Change Detection Using Variable Segments for Video Indexing

King Yiu Tam, Jose Lay, and David Levy

The University of Sydney, School of Electrical and Information Engineering,
Sydney, Australia
`{evantam, jlay, dlevy}@ee.usyd.edu.au`

Abstract. Video indexing based on shots obtained by visual features is useful for content-based video browsing but has more limited success in facilitating semantic search of videos. Meanwhile, recent developments in speech recognition allow the option of surpassing many difficulties associated with the detections of semantic meanings over visual features by operating directly on the verbal content. The use of language based indexing inspires a new video segmentation technique based on speaker change detection. This paper deals with the improvement of existing speaker change detectors by introducing an extra preprocessing step which aligns the audio features with syllables. We investigate the benefits of such synchronization and propose a variable presegmentation scheme that utilizes both magnitude and frequency information to attain such alignment. The experimental results show that the quality of the extracted audio feature is improved, resulting in a better recall rate.

Keywords: Variable Segments, Speaker Change Detection, MFCC, BIC.

1 Introduction

A crucial step in video indexing is the division of a video clip into some meaningful segments by way of syntactic or semantic organization. A similar issue is the indexing of written works. The latter is comparably easier as languages have inherent structures comprising paragraphs, sentences and words. In contrast, the division of a video is somewhat less natural due to its continuous nature. A typical solution is to adopt the cinematic shot segmentation technique where a shot is defined as a single continuous camera recording. The concept of shots as index units is also operationally compelling as shot segmentation can be obtained quite accurately by using perceptual feature based shot boundary detection techniques. These approaches normally use some perceptual features such as colour histograms, motion vectors, etc. to detect where an abrupt change occurs between consecutive frames in a video, signaling the onset of a different camera shot. Sensibly, shot segmentation could be used to generate thumbnails of scenes which are useful for facilitating non-linear browsing of a video by its visual contents. However, as the approach has its root in syntactical visual differences, the use of shots for semantic query such as the indexing of news videos is rather unsatisfactory. A problem is that a shot may not coextend a semantic unit. For

example a single scene may have cuts of shots from multiple viewing angles. An illustration is given in Fig. 1. Fig. 1a shows a short action sequence from a James Bond movie, with the vertical markers showing the start and end of the actions. Similarly Fig. 1b shows the speech associated with the action, again, the markers marks the temporal location of the dialogues. The result from a typical shot detector is shown in Fig. 1c, the first 4 shot changes are due to scenes or camera angles while the last one is due to the gun flashes from Bond's firearm. Intuitively, if these individual shots are automatically or manually annotated, then natural language based search of the video by content also becomes feasible.

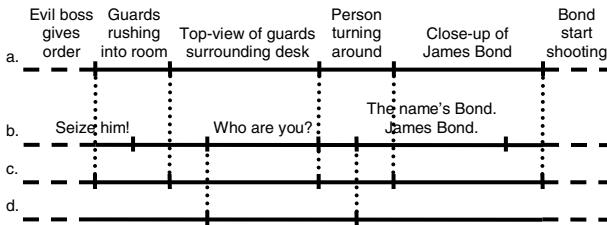


Fig. 1. Typical scene from a James Bond movie (a) Action sequence (b) Actors' speech (c) Result from shot detection (d) Result from speaker change detection

Sensibly, for videos featuring typical conversational contents, a workable alternative is to automatically annotate a video by using the verbal text content in the video. Consequently, we are interested in adopting a language-based video segmentation through speaker changes (Fig. 1d) where a video index unit would co-extend a continuous sequence of verbal utterances made by a speaker. For example, this would allow a segment that begins with the scene where the evil boss gave an order to his guards to seize Bond and terminates and inclusive of the scene where the guards are rushing into the room and surrounding Bond.

Clearly, shot-based indexing and language-based indexing deal with different aspects of video content and are complementary in nature. For example, the shot-based indexing has its aim to allow query such as: “find those shots showing close-up scenes of Bonds”, where the query would be operated by a pool of visual features working to detect the occurrence of close-up objects classified as Bond; the language-based indexing aims to support query rooted in verbal content such as: “find those segments where the boss calls for the capturing of Bond”.

Furthermore, as shot-based segmentation has been well studied and developed, we are interested in improving the language-based video indexing solution through better speaker change detection techniques. Hence this paper will further focus on the improvement of speaker change detection techniques for the purpose of video indexing.

2 Proposed Method

Speaker change detection or speaker segmentation deals with the detection of time indices in a conversational audio stream where utterances of unique speakers begin; there are thus 3 general scenarios to be considered:

- A change point that takes place after a long silence or at the beginning of the audio stream where a person starts speaking.
- A change point where a speaker starts speaking after the other speaker stops speaking.
- An overlapped change where a speaker starts speaking before the other speaker stops.

The problem is rather straightforward if speaker identity is known a priori such as speaker models can be trained and used for the detection. Unfortunately speaker identity is usually unknown in many cases, for example, consider a news clip; hence speaker models need to be estimated from within the data. Previous studies have looked at ways of improving model estimation through the use of Vector Quantization [1], Gaussian Mixture Model [1-3] and Hidden Markov Model [2]. However most of them have overlooked the feature extraction step, which plays no less importance in obtaining accurate detection. This paper revisits the feature extraction step and attempts to improve model estimation by synchronizing the feature segmentation with syllable boundaries.

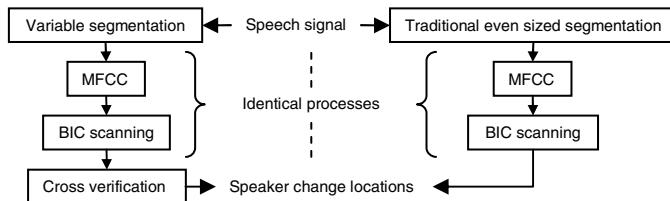


Fig. 2. Block diagram of the proposed speaker change detection method (*left*) and the traditional method (*right*)

A speaker change detection algorithm usually begins with the segmentation of the speech signal into smaller units [1-3]. The smaller segments are then seen as small neighborhoods from which feature vectors are extracted. Ideally the segments size should be chosen such that its multiples will coincide as closely as possible with the syllables in the signal so that each syllable can be separated exactly from its neighbors. Intuitively, when the segments are perfectly aligned with the syllables, the quality of the feature vectors will be higher as opposed to when there is no alignment. In practice, the signal is normally segmented into even sized segments only without alignment under the assumption that the tempo of the speech will cluster around some mean value even though it is variable. Although speech clearly is rhythmic, the duration of the syllables which form the speech is not. Moreover there will be offset due to intonation of the speaker or the rhythm will change when there is a speaker change.

This work builds upon earlier works [1-3] in using Bayesian Information Criterion (BIC) for speaker change detection. More specifically, we propose to replace the even sized segments with variable segments before Cepstrum features are computed. Fig. 2 shows the general block diagram of the proposed method comprising three main parts: variable segmentation, BIC scanning and cross verification.

Let's imagine for a moment that there is a perfect syllable detector which is able to detect the start of every syllable. Since a speaker change cannot occur in the middle of a syllable, speaker change location is also a subset of the syllable detector output. Therefore if the BIC scanning algorithm is forced to operate on the locations returned from the syllable detector, then the result will be significantly more accurate compare to a blind search because the search range is limited to discrete positions instead of a range. An analogy to this is the task of locating the start of every sentence in this text. The reader can accurately determine the start of every sentence only because punctuation marks are provided. An important point is that only full stops signal the limit of sentences and that it is a subset of all punctuation marks. In this case the syllable detector output to a speech signal is the same as the punctuation marks to this text and that speaker change locations is the same as full stops. The reader then examines only punctuation mark locations to decide if it is the start of a sentence. If the same task is carried out without punctuations then low precision, low recall and high false alarm is expected even for experienced readers. Before moving on to deeper discussions, here are some terminologies:

- Subsegment – Before a speech signal can be analyzed, it must be split into chunks, where each chunk is consider as a small neighborhood from which a single feature vector can be used to describe that neighborhood.
- Subsegment Size – The length of speech consider as a subsegment. It is usually a common factor of the different syllable durations also it is a width which is large enough to capture the frequency range of interest.
- Force Alignment Points – Output from the varseg boundary detector in Sect. 2.1. The start of every syllable is a subset of this output.
- Subsegment Points – The boundary between neighboring subsegments after they are spaced out from force alignment points. The distance between two subsegment points is exactly one subsegment size.
- Variable Segment – Can occur in two ways: 1. Consider a length of speech of a single syllable, the variable segment associate with this syllable is the residue after division by the subsegment size. 2. A sudden change, possibly noise in the speech signal that has duration of less than the subsegment size.
- Region – Area of similar characteristics, for example all the data between the start and the end of a syllable. The data within the duration of noise, however small that duration, is considered a single region. Likewise, all the data from the start to finish of silence is considered as a region.

Fig. 3a shows the ideal noise-free signal and the actual signal Fig. 3b. If conventional method is used, the signal will be naively divided (Fig. 3b). Therefore the result often includes subsegments which span across region boundaries, namely silent + noise, silent + syllable and syllable + noise, which will compromise the overall accuracy of the system. In pursuit of a better alignment, this paper proposes a variable segmentation scheme to replace the traditional method in which most subsegments are aligned to the beginning of syllables (Fig. 3c). Provided that enough of the subsegments are aligned, the probability of detection at the correct positions is increased. The variable segments serve as ‘wiggling room’ such that other subsegments do not span into neighboring regions, these are shown as hashed areas (Fig. 3c). Fig. 3d shows the result after the variable segments are dropped; the

subsegments in Fig. 3c and Fig. 3d are numbered to show the ‘before’ and ‘after’ locations. Attaining this alignment should in theory improve the quality of the extracted feature vector and in turn the overall accuracy of speaker change detection.

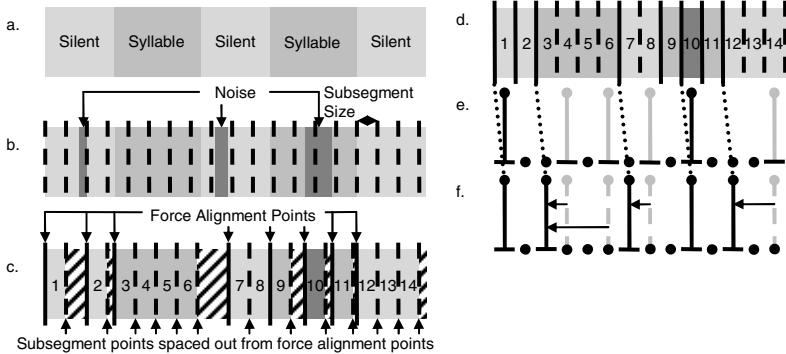


Fig. 3. Segmentation of a speech signal (a) Ideal signal (b) Traditional method (c) Proposed method (d) After variable segments are discarded (e & f) Cross verification

2.1 The Variable Segment Feature

The boundary detector is the core component of the variable segmentation scheme, this algorithm not only detects the beginning of syllables, it also detects all the transition location between silences, noise and syllables. The detection of these boundaries, also known as force alignment points, allows the clip to be separated into silence, noise or syllable regions. These force alignment points are shown as solid lines in Fig. 3c. The subsegment points shown as dotted lines marks the edge of each subsegment spaced out from force alignment points. The following is an example showing the calculation of the varseg feature within the boundary detector. Fig. 5 illustrates the variable segmentation of a part of a sentence “each year thousands of children are” taken from a current event TV program. The boundaries detection steps are explained below:

1. The analytical signal $S_a(t)$ is extracted via Hilbert Transform [4] in the form:

$$S_a(t) = a(t)e^{j\phi(t)} \quad (1)$$

Where magnitude $a(t)$ and phase $\phi(t)$ is given by:

$$a(t) = |S_a(t)| \quad (2)$$

$$\phi(t) = \angle S_a(t) \quad (3)$$

$a(t)$ and $\phi(t)$ are illustrated in Fig. 5d and Fig. 5f respectively.

2. The change in magnitude $a'(t)$ and instantaneous frequency (IF) $\phi'(t)$ is given by the first derivative of $a(t)$ and $\phi(t)$ respectively:

$$a'(t) = \frac{da(t)}{dt} \quad (4)$$

$$\phi'(t) = \frac{d\phi(t)}{dt}. \quad (5)$$

The change in frequency $\phi''(t)$ is given by taking the first derivative of IF:

$$\phi''(t) = \frac{d\phi'(t)}{dt}. \quad (6)$$

$a'(t)$, $\phi'(t)$ and $\phi''(t)$ are respectively illustrated in Fig. 5e, Fig. 5g and Fig. 5i. Note that if the IF is unwrapped (Fig. 5h), a waveform somewhat digital like will result, it is found that the ‘steps’ of this waveform correspond to abrupt changes such as syllable, word or speaker changes.

3. In Fig. 5j the analytical signal is reorganized into a more useful form by taking the product of change in magnitude and change in frequency. The varseg feature v is given by:

$$v = a'(t) \times \phi''(t). \quad (7)$$

4. Due to the fast changing nature of speech, the transition boundaries between neighboring regions are not always easy to detect. Therefore a threshold based on standard deviation of the varseg feature v is used, by tweaking this threshold value, it is possible to select the minimum change in magnitude of the varseg feature required to be classified as the start of a new region. Fig. 5k shows the result after thresholding at 1 standard deviation. At this stage, there seems to be many false detections, however this is intentional because they serve as force alignment points. The result at this stage is identical to the force alignment points in Fig. 3c. As long as all the boundaries corresponding to syllable change are included in this result, then the feature would have achieved its intended purpose. After some filtering based on maximum neighboring distances the result is shown in Fig. 5c. The Comparison between Fig. 5a. and Fig. 5c proofs that the starting location of every syllable is included in the boundary detection result.
5. Finally to complete the variable segmentation process from step 4, a predefined subsegment size is used to space out the other subsegments from the force alignment points as shown in Fig. 3c. The residues identical to the hashed areas in Fig. 3c are then dropped.

2.2 BIC Scanning Algorithm

After initial pre-segmentation, MFCC vector is extracted for each subsegment, these are further processed by the BIC scanning algorithm.

Using Fig. 4 to aid our discussion, the scanning algorithm is performed as follows:

1. Consider a speech signal with a total length of 2 add size in Fig. 4a. The algorithm will initialize with a total length of 1 add size (the left half of Fig. 4a), this portion of the feature is called the scanning range.
2. The scanning range is further divided into two parts; the divider can be moved from left to right in step size specified by resolution. There are limits of 1 pad size at both ends of the scanning range to ensure the size of each half is of reasonable length and not go to zero. This is shown in Fig. 4b.

3. For each iteration, the divider is moved one step to the right and BIC value is calculated for that position. After the divider got to the right end, if the maximum overall BIC value is greater than zero then a speaker change has been detected at that maximum position.
4. If a candidate has been found, the algorithm reinitialize as if the speech signal begins at that position. A buffer of history is added at the beginning (left end). The new scanning range is then of length buffer size + add size as shown in Fig. 4c.
5. However if no candidate is found then the new scanning range is just the previous plus a length of add size as shown in Fig. 4d. Repeat from step 2 to 5 until the end of the signal is reached.

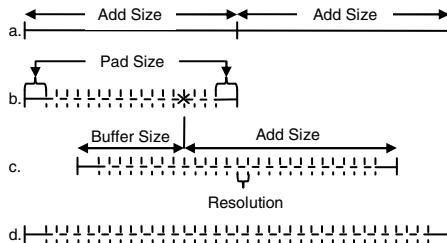


Fig. 4. BIC scanning algorithm

2.3 Cross Verification

The next step is to combine the result of the boundary detector from the initial presegmentation step with the output of the BIC scanning algorithm. As illustrated previously, the result from the boundary detector (Sect. 2.1 Step 4) already contains all the possible speaker change locations. Therefore it makes sense to realign the BIC scanning result using the boundary detector output. This behavior is illustrated in Fig. 3d, Fig. 3e and Fig. 3f. Since the absolute positions are stored locally within the subsegments, even if the variable segments are dropped, the actual positions do not change; this relationship is shown by the numbers in Fig. 3c and Fig. 3d. After the MFCC from each sub-segment is extracted from Fig. 3d, they are processed by the BIC scanning algorithm. These originally detected positions contain errors because the original signal contains noise (Fig. 3e). The false detections are shown in gray while the correctly detected positions are in black. Although the boundary detector output and the BIC scanning result both contain errors, the combination of the two will still have less error because the parts contain different information. This cross verification step can be thought of as the BIC scanning results latching onto the nearest boundary detector result to the left (Fig. 3f). The small dotted lines linking Fig. 3d and Fig. 3f shows how the BIC scanning result latch onto forced alignment points.

3 Experimental Result

This section presents the comparison between the proposed method and traditional method in combination with BIC for the application of speaker change detection. First the signal was presegmented into subsegments using the steps discussed in Sect. 2.1.

In step 4, the signal was thresholded at 1 standard deviation. From [5] it is known that for normal speech, the mean syllable duration for English vocabulary is 125ms and that the pause between syllables is about 100ms. Therefore it is logical to set the sub-segment size to be a common factor between 125 and 100. Experimentally 5ms is found to be the best value. The next step is MFCC extraction follow by the BIC scanning algorithm. From experience the add size involved in the scanning algorithm has the largest impact on precision and record rate, hence this is calculated from known English statistics. From [5] the mean syllable width was 125ms and the pause between syllables is 100ms. From [6] it is known that the mean syllable per word in the English vocabulary is 1.5. From [7] it is known that the average number of words per sentence in written works range from 17 to 20; therefore the average is 18.5. From the above statistics, the mean sentence length is:

$$(125\text{ms}+100\text{ms}) \times 1.5 \times 18.5 = 6244\text{ms}. \quad (8)$$

Due to the subsegment size of 5ms determined previously, the optimum add size is:

$$6244\text{ms} \div 5\text{ms} = 1249. \quad (9)$$

This is basically the number of 5ms subsegment required to span the average width of one sentence. The detected speaker change points are then subject to a cross verification step as explained in Sect. 2.3. The performance of the proposed method will be discussed in the next section.

4 Discussion

As shown in Fig. 2, the MFCC calculation and BIC algorithm processes are identical for the traditional method and the proposed method. Both methods uses the first 13 MFCC coefficients, the only difference is the added processes associated with varseg. For the traditional method, large amount of samples are normally skipped according to a predefined frame rate before MFCC calculation. This is a well known procedure to remedy the weak classification power of BIC by increasing the standard deviation of the MFCC vectors through skipping. A major drawback of this mechanism is that the offset between the detected location and the actual location can be quite large because the actual speaker change may correspond to the samples which have been skipped. Moreover, the algorithm may fail to detect the speaker change if the change is subtle; the information required for differentiating the speakers maybe contained in the skipped frames. Instead of skipping samples, the proposed method relies on the sample/syllable synchronization to boost the strength of the extracted MFCC vectors. Also the result is further improved through cross verification between the result from boundary detector and BIC scanning algorithm. Since the proposed method does not rely on skipping to increase precision, it is able to respond more frequently and therefore able to pick up more speaker changes. The comparison between varseg and the traditional method is show in Table 1, the comparison is benchmarked at +/- 0.5s and +/- 1s tolerance. For example, +/- 1s tolerance means that the distance between the detected point and the true location must be less than or equal to 1 second for it to count towards the recall and precision. The clips are divided into three types:

- Studio – Low noise, near ideal recording environment.
- Hybrid – Contain cuts from both indoor and outdoor environments.
- Outdoor – High background noise environment.

There are two sets of results for the proposed method; the theoretical and tuned configurations. These configurations refer to the different add size settings. The add size for the theoretical configuration was calculated in Sect. 3 Eq. 9, while the tuned version was manually tweaked to Clip 3, 8, 9, 12 and 14 because the traditional

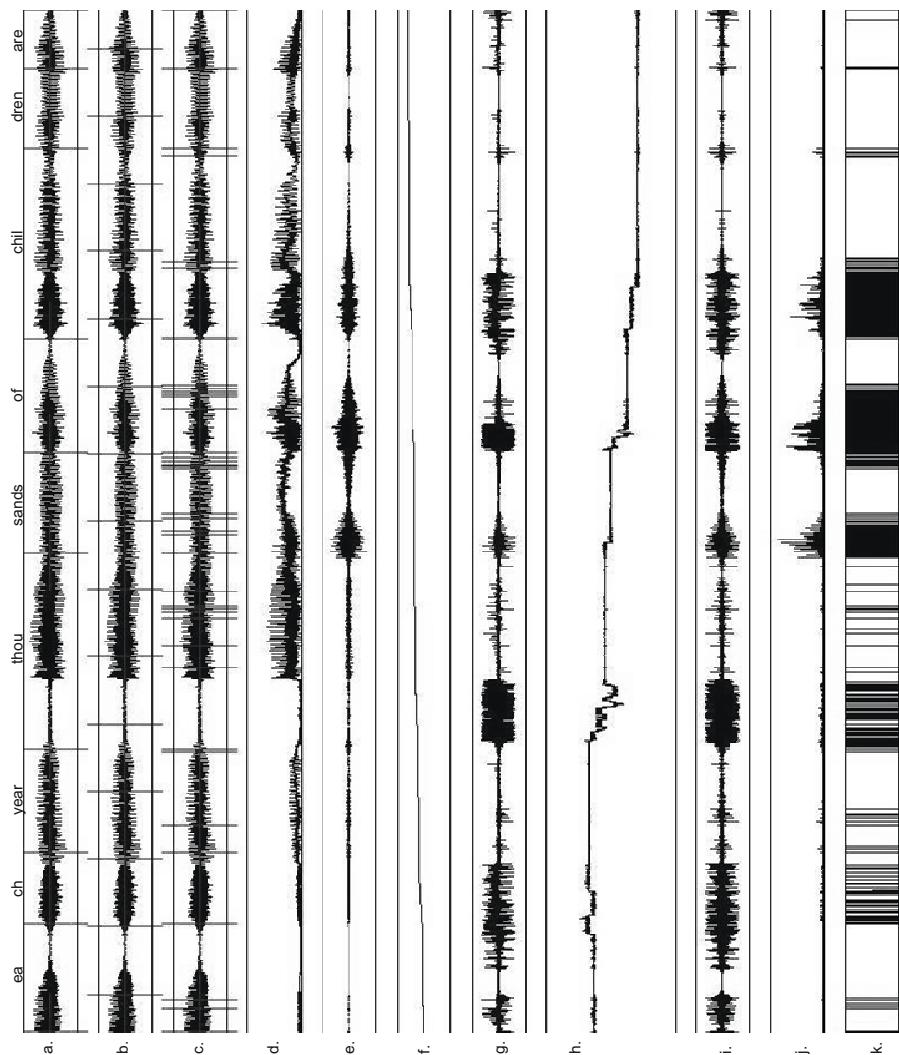


Fig. 5. Preprocessing steps of proposed method and comparison with traditional method

method performed poorly on them; the average of the best setting for these clips was used as the add size. From Table 1 it can be seen that at +/- 0.5s tolerance, the proposed method under the tuned configuration has roughly 10% higher recall compare to the traditional method at almost the same level of precision. When the test was carried out at +/- 1s tolerance, the proposed method suffer roughly 10% lower precision but maintained the 10% lead in recall. The standard deviation for the proposed method is lower compare to the traditional method in almost all cases, indicating that the performance of the proposed method is more consistent. Assuming that the tuned configuration is the ideal setting, due to the similarity of the results between the theoretical and the tuned configuration, we suspect that there is a close relationship between recall/precision and the matching between the add size and the average number of words per sentence (AWS) of the clip. To investigate this further, we propose an additional step called Sentence Queue Optimization (SQO) for cases when the transcript is available, as in the case for documentaries and news broadcast. Specifically the SQO Add Size for each clip can be calculated directly from the transcript using Eq. 8:

$$\text{SQO Add Size} = (125 + 100) \times 1.5 \times \text{AWS}, \text{ AWS} = \frac{\text{No. spaces}}{\text{No. full stops}}. \quad (10)$$

Table 1. Comparison between varseg and traditional method

Duration (Minute)	Tolerance +/- 0.5s (All units in %)								Tolerance +/- 1.0s (All units in %)									
	Varseg Theoretical		Varseg Tuned		Traditional		Varseg Theoretical		Varseg Tuned		Traditional		Varseg Theoretical		Varseg Tuned		Traditional	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
Studio	61	49	62	55	44	44	77	70	72	64	60	51	51	51	51	89	89	
	50	48	Avg 48	56	48	49	63	60	67	67	68	67	41	41	41	86	86	
	31	38	Avg 49	56	41	41	33	44	54	56	59	59	65	65	65	65	65	
	48	42	Avg 50	52	44	44	19	20	54	56	59	59	56	56	56	81	81	
	16:01	48	Avg 50	52	44	44	19	20	54	56	59	59	56	56	56	81	81	
	07:56	55	Avg 50	55	50	50	17	55	61	65	65	65	53	53	53	84	84	
Hybrid	67	60	Avg 50	55	50	50	17	50	50	64	64	64	55	55	55	85	85	
	03:34	86	Avg 50	55	50	50	17	50	50	78	78	78	60	60	60	89	89	
	02:48	60	Avg 50	55	50	50	10	55	64	64	64	64	50	50	50	81	81	
	02:47	55	Avg 50	55	50	50	10	55	60	60	60	60	50	50	50	81	81	
	09:35	47	42	42	42	42	32	41	62	55	65	58	59	59	59	74	74	
	05:31	41	37	41	32	32	59	71	59	53	59	45	71	71	71	86	86	
Outdoor	32	44	44	36	47	32	50	50	69	69	75	71	71	71	71	86	86	
	04:06	46	46	53	53	53	35	64	60	53	50	55	55	55	55	86	86	
	03:43	35	Avg 46	50	45	45	30	50	53	53	53	53	53	53	53	86	86	
	03:38	60	Avg 46	50	42	42	30	50	53	53	53	53	53	53	53	86	86	
	03:33	64	Avg 46	50	42	42	30	50	53	53	53	53	53	53	53	86	86	
	03:31	73	57	82	50	50	45	42	73	73	73	73	73	73	73	73	73	
	03:24	57	53	64	41	64	65	79	79	57	50	50	47	47	47	73	73	
	Avg	50	53	52	53	40	54	61	66	64	64	64	54	54	54	73	73	
	STD	13	14	12	13	14	18	10	14	10	14	10	14	14	14	15	15	

Table 2. Comparison of the effect of SQO in combination with varseg

Duration (Minute)	Tolerance +/- 1s		Theoretical		with SQO	
	Recall	Precision	Recall	Precision	Recall	Precision
01:49	25%	29%	38%	33%		
02:50	40%	67%	47%	78%		
02:58	50%	38%	50%	43%		
03:50	41%	69%	44%	71%		
04:07	29%	60%	33%	58%		
04:21	36%	56%	46%	62%		
04:57	80%	57%	80%	50%		
16:50	40%	77%	43%	72%		
Avg	43%	56%	48%	58%		
Std	17%	16%	14%	15%		

Table 2 shows the additional effect from SQO in combination with the proposed method. The clips used here are completely different from the ones used previously for Table 1 due to the lack of transcripts; however the result is still a valid indicator of what SQO can do. As shown in the table, the recall rates of most clips have improved with a mean of 5% compare to the results in which the add size was derive from the theoretical approach.

5 Conclusion and Future Work

We have shown the proposed method is able to achieve higher recall and consistency at a precision similar to the traditional method. We have also shown that it is possible for the proposed method to be automatically tuned when a transcript is available. As this was initially developed as part of a video indexing and searching system, the next step is to combine it with an automatic speech recognition system to automatically annotate video clips.

References

1. Mori, K., Nakagawa, S.: Speaker Change Detection and Speaker Clustering using VQ Distortion for Broadcast News Speech Recognition. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2001), Salt Lake City UT USA, vol. 1, pp. 413–416 (2001)
2. Radhakrishnan, R., Xiong, Z., Divakaran, A., Raj, B.: Investigation on Effectiveness of Mid-level Feature Representation for Semantic Boundary Detection in News Video. In: Internet Multimedia Management Systems, Conference, Orlando FL, vol. 5242(4), pp. 74–80 (2003)
3. Lu, L., Zhang, H.: Real-Time Unsupervised Speaker Change Detection. In: Proc. 16th International Conference on Pattern Recognition 2002, vol. 2, pp. 358–361 (2002)
4. Paliwal, K., Atal, B.: Frequency Related Representation of Speech. In: Proc. European Conf. Speech Communication and Technology, EUROSPEECH 2003, Geneva Switzerland, pp. 65–68 (2003)
5. Speech Frequency Components,
<http://www.smeter.net/daily-facts/4/fact2.php>
6. Yaruss, J.: Converting Between Word and Syllable Counts in Children’s Conversational Speech Samples. Journal of Fluency Disorders 25(4), 305–316 (2000)
7. Document Readability on My Writer Tools,
<http://www.mywritertools.com/lightenup.asp>

Correlated PLSA for Image Clustering

Peng Li^{1,2}, Jian Cheng^{1,2}, Zechao Li^{1,2}, and Hanqing Lu^{1,2}

¹ National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing, 100190, P.R. China

² China-Singapore Institute of Digital Media, Singapore 119613, Singapore
`{pli,jcheng,zcli,luhq}@nlpr.ia.ac.cn`

Abstract. Probabilistic Latent Semantic Analysis (PLSA) has become a popular topic model for image clustering. However, the traditional PLSA method considers each image (document) independently, which would often be conflict with the real occasion. In this paper, we presents an improved PLSA model, named Correlated Probabilistic Latent Semantic Analysis (C-PLSA). Different from PLSA, the topics of the given image are modeled by the images that are related to it. In our method, each image is represented by bag-of-visual-words. With this representation, we calculate the cosine similarity between each pair of images to capture their correlations. Then we use our C-PLSA model to generate K latent topics and Expectation Maximization (EM) algorithm is utilized for parameter estimation. Based on the latent topics, image clustering is carried out according to the estimated conditional probabilities. Extensive experiments are conducted on the publicly available database. The comparison results show that our approach is superior to the traditional PLSA for image clustering.

Keywords: Correlated PLSA, topic model, image clustering.

1 Introduction

Image clustering is the process of grouping similar images together. It is a basic problem in many applications such as image annotation, object recognition, image retrieval. Although it has been studied for many years, it is still a challenging problem in multimedia and computer vision communities.

There are many widely used clustering methods for image clustering, e.g K-means, Gaussian Mixture Model (GMM), etc. However, most clustering methods (eg. K-means) perform clustering intuitively by calculating the distance between the data points and the cluster centers, which often lead to poor clustering results. In recent years, topic models such as Latent Semantic Analysis (LSA) [3] and Probabilistic Latent Semantic Analysis (PLSA) [5] have become popular tools to handle the problem. For these topic models, the images are modeled by some latent topics, which are semantic middle level layers upon the low level features and more discriminative.

The topic models were originally proposed to handle text corpus problems. In [3], LSA used Singular Value Decomposition (SVD) of the word-document matrix to identify a latent semantic space. However, LSA has a number of deficits due to its unsatisfactory statistical formulation [5]. In order to overcome this problem, Hofmann proposed a generative probabilistic model named Probabilistic Latent Semantic

Analysis (PLSA) in [5]. PLSA models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of "topics". Due to the success of the topic models in text analysis, they have been introduced into the field of computer vision and multimedia to solve various problems [1, 8, 9, 10, 11, 12, 15]. PLSA is used for image classification in [1, 8]. Shah-hosseini A. et al. [12] did semantic image retrieval based on the PLSA model. Zhang R.F. et al. [15] used PLSA for hidden concept discovery by segmenting the images into regions. In [9], a latent space is constructed with the PLSA model and image annotation is done based on the latent space. Peng Y.X. et al. [10] constructed an audio vocabulary and proposed an audio PLSA model for semantic concept annotation.

However, there is a problem with the PLSA model. It doesn't consider the image correlations when estimating the parameters, which often leads to inaccurate latent topics. For example, in image clustering task, some similar images that should be in the same cluster always have different topic distributions, which leads to bad clustering results. Actually, there exists much latent semantic correlation among images or image regions. Therefore, it is natural that the correlations between images should be incorporated into the topic model in order to derive more accurate latent topics. Inspired by [4], we propose a Correlated Probabilistic Latent Semantic Analysis (C-PLSA) model in this paper. In our model, we introduce a correlation layer between the images and the latent topics to incorporate the image correlations. We apply the C-PLSA model to image clustering and the experiment results show that our model can get very promising performance.

The rest of this paper is organized as follows: in Section 2, we give a brief review of the PLSA model. Section 3 gives the detail of our C-PLSA model. Experiment results are presented in Section 4. Finally, we conclude our paper in Section 5.

2 The PLSA Model

The PLSA model was originally developed for topic discovery in a text corpus, where each document is represented by its word frequency. The core of PLSA model is to map high dimensional word distribution vector of a document to a lower dimensional topic vector. Therefore, PLSA introduces a latent topic variable $z_k \in \{z_1, \dots, z_K\}$ between the document $d_i \in \{d_1, \dots, d_N\}$ and the word $w_j \in \{w_1, \dots, w_M\}$. Then the PLSA model is given by the following generative scheme:

1. select a document d_i with probability $P(d_i)$,
2. pick a latent topic z_k with probability $P(z_k | d_i)$,
3. generate a word w_j with probability $P(w_j | z_k)$.

As a result one obtain an observation pair (d_i, w_j) while the latent topic variable z_k is discarded. This generative model can be expressed by the following probabilistic model:

$$P(w_j, d_i) = P(d_i)P(w_j | d_i), \quad (1)$$

$$P(w_j | d_i) = \sum_{k=1}^K P(w_j | z_k)P(z_k | d_i). \quad (2)$$

The model is graphically in Fig. 1.

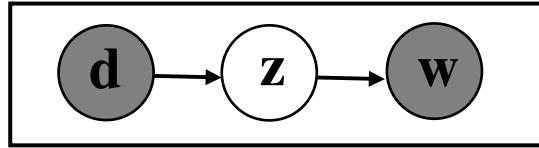


Fig. 1. The PLSA model

We learn the unobservable probability distribution $P(z_k | d_i)$ and $P(w_j | z_k)$ from the complete dataset using expectation maximization (EM) algorithm [2]. The log-likelihood of the complete dataset is:

$$\begin{aligned} L &= \sum_{i=1}^N \sum_{j=1}^M n(d_i, w_j) \log P(d_i, w_j) \\ &\propto \sum_{i=1}^N \sum_{j=1}^M n(d_i, w_j) \log \sum_{k=1}^K P(w_j | z_k) P(z_k | d_i) \end{aligned} \quad (3)$$

where $n(d_i, w_j)$ is the number of occurrences of word w_j in document d_i . The E-step is given by

$$P(z_k | d_i, w_j) = \frac{P(w_j | z_k) P(z_k | d_i)}{\sum_{l=1}^K P(w_j | z_l) P(z_l | d_i)}, \quad (4)$$

and M-step is given by

$$P(w_j | z_k) = \frac{\sum_{i=1}^N n(d_i, w_j) P(z_k | d_i, w_j)}{\sum_{i=1}^N \sum_{j=1}^M n(d_i, w_j) P(z_k | d_i, w_j)}, \quad (5)$$

$$P(z_k | d_i) = \frac{\sum_{j=1}^M n(d_i, w_j) P(z_k | d_i, w_j)}{\sum_{j=1}^M n(d_i, w_j)}. \quad (6)$$

Iteratively perform E-step and M-step until the probability values are stable.

3 Our Correlated PLSA Model

3.1 Overview

Although the PLSA model was originally developed for topic discovery in a text corpus, it has been introduced into multimedia field due to its success in recent years, for example, image annotation, object recognition, etc. When applied to images, each image represents a single document and the words can be replaced by visual words,

image regions, etc. However, there is a problem with the PLSA model: it doesn't consider the image correlations when estimating the parameters. In order to derive more accurate latent topics, we propose an improved Correlated PLSA (C-PLSA) model to address the correlations between the images in the dataset.

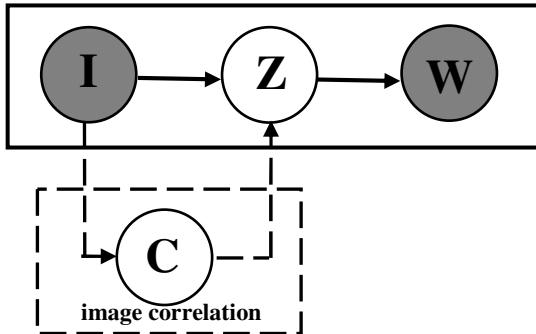


Fig. 2. The C-PLSA model

We depict an overview of our C-PLSA model in Fig.2. Given the image $I_i \in \{I_1, \dots, I_N\}$, the visual word $W_j \in \{W_1, \dots, W_M\}$ and the latent topic $Z_k \in \{Z_1, \dots, Z_K\}$, we adopt the same generative scheme as that of PLSA, which is shown in the solid box in Fig.2:

1. select an image I_i with probability $P(I_i)$,
2. pick a latent topic Z_k with probability $P(Z_k | I_i)$,
3. generate a visual word W_j with probability $P(W_j | Z_k)$.

In addition, we introduce a new correlation layer in our model such that the topic distributions of the given image can be updated by that of the images similar to it. The image correlations are parameterized by the image correlation matrix C as is shown in the dashed box in Fig.2. As we have incorporated the image correlations into the PLSA model, the related images can have similar topic distributions and we will get more accurate latent topics than the PLSA model.

3.2 Bag-of-Visual-Words Representation and Image Correlations

When we use the C-PLSA model, the bag-of-words image representation has to be generated first. Here the generation of bag-of-visual-words consists three major steps. First, Difference of Gaussian (DoG) filter is applied on the images to detect a set of key points and scales respectively. Then, we compute the Scale Invariant Feature Transform (SIFT) [7] over the local region defined by the key point and scale. Finally, we perform vector quantization on SIFT region descriptors to construct the visual vocabulary by exploiting the hierarchical k-means clustering methods. Then we can get the word-image matrix (see Fig.3). Each row in the matrix represents an image and $n(I_i, W_j)$ specifies the number of times the visual word W_j occurred in image I_i .

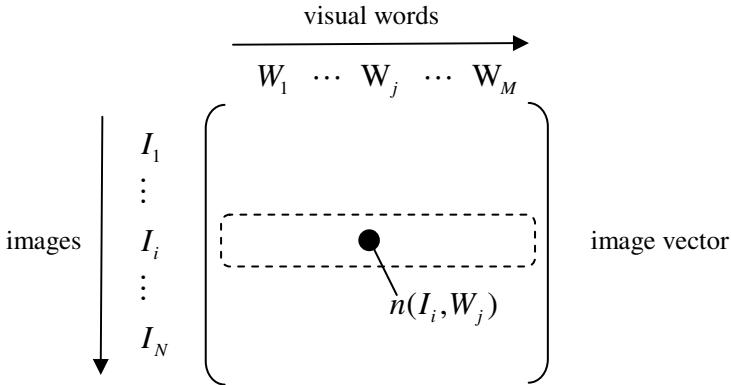


Fig. 3. Word-image matrix

With the bag-of-visual-words vector representation introduced above, we compute the image correlation matrix C by cosine similarity. For each pair of images in the dataset, we first compute their cosine similarity as follows:

$$\text{Sim}_{ih} = \frac{\vec{I}_i \cdot \vec{I}_h}{\|\vec{I}_i\| \cdot \|\vec{I}_h\|}, \quad (7)$$

where \vec{I}_i is the i -th image and represented by the i -th row in the word-image matrix. Then we can get a similarity matrix S where $S_{ih} = \text{Sim}_{ih}$. After we get the similarity matrix S , we only keep Q nearest neighbors for each image. In other words, only the top Q values in each row of S are kept and the others are set to zero. At last, we get the image correlation matrix C by normalizing the matrix S such that its row add up to 1.

$$C_{ih} = \frac{S_{ih}}{\sum_{h=1}^N S_{ih}} \quad (8)$$

Therefore, each element of C can be considered as the conditional probability $P(I_h | I_i)$ and the topic distribution of a given image can be updated by the topic distributions of the images that are related to the given image as follows:

$$P(Z_k | I_i) = \sum_{h=1}^N P(Z_k | I_h) P(I_h | I_i). \quad (9)$$

Since we have introduced a correlation layer between the images and the latent topics in our C-PLSA model, we can derive more accurate topic distributions than the traditional PLSA model.

3.3 Parameter Estimating

Following the maximum likelihood principle, we estimate the parameters $P(Z_k | I_i)$ and $P(W_j | Z_k)$ by maximizing the log-likelihood function:

$$L = \sum_{i=1}^N \sum_{j=1}^M n(I_i, W_j) \log \sum_{k=1}^K P(W_j | Z_k) P(Z_k | I_i), \quad (10)$$

and EM algorithm can be used to estimate the parameters. In order to incorporate the image correlations, we renew the probability $P(Z_k | I_i)$ by equation (9) at each end run of the M-step, thus resulting in a variation of EM algorithm through the following expectation (E-step) and maximization (M-step) solution.

The E-step is given by

$$P(Z_k | I_i, W_j) = \frac{P(W_j | Z_k) \overline{P(Z_k | I_i)}}{\sum_{l=1}^K P(W_j | Z_l) \overline{P(Z_l | I_i)}}, \quad (11)$$

and the M-step is given by

$$P(W_j | Z_k) = \frac{\sum_{i=1}^N n(I_i, W_j) P(Z_k | I_i, W_j)}{\sum_{i=1}^N \sum_{j=1}^M n(I_i, W_j) P(Z_k | I_i, W_j)}, \quad (12)$$

$$P(Z_k | I_i) = \frac{\sum_{j=1}^M n(I_i, W_j) P(Z_k | I_i, W_j)}{\sum_{j=1}^M n(I_i, W_j)}, \quad (13)$$

$$\overline{P(Z_k | I_i)} = \sum_{h=1}^N P(Z_k | I_h) P(I_h | I_i). \quad (14)$$

Iteratively perform E-step and M-step until the probability values are stable.

4 Experimental Evaluations

In this section, we evaluate our C-PLSA model by comparing it with the traditional PLSA and K-means on the Caltech-101 Object Categories [13]. Some category names and randomly selected sample images are shown in Fig.4. Each object category contains about 40 to 800 images and a unique label has been assigned to each image to indicate which category it belongs to, which serves as the ground truth in the performance studies. We first compute the word-image matrix by extracting SIFT features as described in Section 3.2. The dimension of the bag-of-visual-words is set to 1000 in the experiment. Then all the clustering methods are performed on the word-image matrix to generate K clusters.

For the topic models, we run the EM algorithm multiple times with random starting points to improve the local maximum of the EM estimates. To make comparison fair, we use the same starting points for PLSA and C-PLSA. The maximum iteration times is set to 150. After representing all the images in terms of latent topic space, each image can be assigned to the most probable latent topic according to the topic distributions $P(Z_k | I_i)$. As respect to K-means, we implement the algorithm on the word-image matrix by computing Euclidean distance between image vectors and the randomly initialized cluster centers until the cluster centers are not changed.



Fig. 4. Some sample images from the Caltech-101 Object Categories

The clustering result is evaluated by comparing the obtained cluster label of each image with that provided by the dataset. The accuracy (AC) [14] is used to measure the clustering performance. Given an image I_i , let r_i and s_i be the obtained cluster label and the label provided by the dataset respectively. The AC is defined as follows:

$$AC = \frac{\sum_{i=1}^N \delta(s_i, map(r_i))}{N} \quad (15)$$

where N is the total number of images and $\delta(x, y)$ is the delta function that equals to 1 if $x = y$ and zero otherwise, and $map(r_i)$ is the permutation mapping function that maps each cluster label r_i to the equivalent label from the dataset. The best mapping function can be found by using Kuhn-Munkres algorithm [5].

The evaluations are conducted for different number of clusters K ranging from 2 to 10. At each run of the test, the images from a selected number K of categories are mixed and provided to the clustering methods. For each given cluster number K , 10 test runs are conducted on different randomly chosen categories, and the final performance scores are obtained by averaging the scores over the 10 test runs.

We first test our C-PLSA model at different numbers of Q and the comparison results are given in Fig.5. As we expected, the accuracy decreases when Q is increasing, because more noise will be introduced into the image correlation matrix. We test different numbers of Q in our experiment and get the best results when Q is around 5.

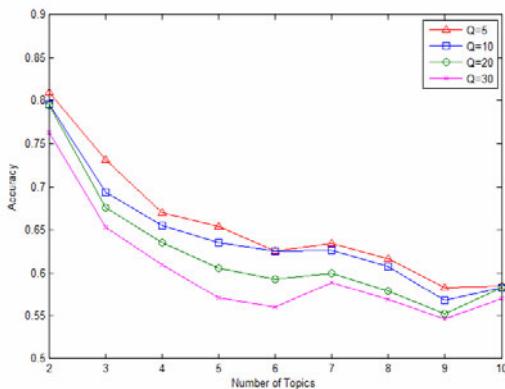


Fig. 5. Accuracy of C-PLSA at different numbers of Q

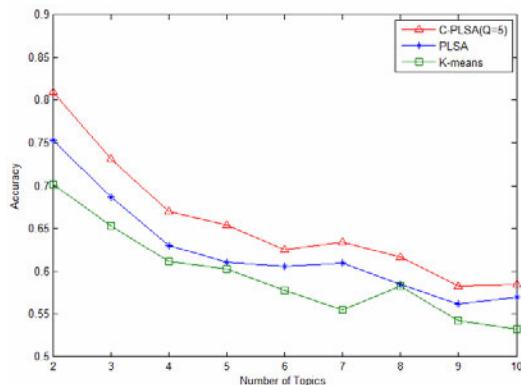


Fig. 6. Accuracy comparison between C-PLSA and other methods

The accuracy comparisons between C-PLSA ($Q=5$) and other methods are reported in Fig.6, which shows that our C-PLSA model outperforms PLSA and traditional K-means in terms of accuracy. It is also in line with our expectation: the correlation information do offer help in deriving more accurate latent topics.

5 Conclusions and Future Work

In this paper, we have presented a novel approach for topic modeling named Correlated Probabilistic Latent Semantic Analysis (C-PLSA). The C-PLSA model introduces a correlation layer between the images and the latent topics, which incorporates the image correlations for topic modeling. As a result, our model can generate more accurate latent topics and have more discriminative power than the traditional PLSA model. The experiment results also show that the image correlations do offer help in the process of topic modeling.

Several questions remain to be investigated in our future work:

1. We consider the image correlations in topic modeling and develop our model based on PLSA. The idea of exploiting image correlations can also be naturally incorporated into other clustering methods, eg., K-means.
2. We compute the image correlations only by bag-of-visual-words features in the paper. More visual features can be combined to get more accurate image correlations.
3. Visual features can't reflect the semantic information of images correctly in many cases. It is very interesting to explore other ways to capture image correlations. For example, web information such as image tags and hyperlink information may be a good way to construct the semantic correlations for web images.

Acknowledgement

This work is partially supported by the National Natural Science Foundation of China (Grant No. 60975010, 60873185, 60833006).

References

1. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via pLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)
2. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from in complete data via the EM algorithm. *Journal of the Royal Statistical Society B* 39, 1–38 (1977)
3. Deerwester, S., Dumais, G.W., Furnas, S.T., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41, 391–407 (1990)
4. Guo, Z., Zhu, S.H., Chi, Y., Zhang, Z.F., Gong, Y.H.: A latent topic model for linked documents. In: Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 720–721. ACM Press, NY (2009)

5. Hoffmann, T.: Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning* 42(1), 177–196 (2001)
6. Lovász, L., Plummer, M.D.: *Matching Theory*. North-Holland, Amsterdam (1986)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal Computer Vision* 60(2), 91–110 (2004)
8. Lu, Z.W., Peng, Y.X., Horace, H.S.Ip.: Image categorization via robust pLSA. *Pattern Recognition Letters* 31(1), 36–43 (2010)
9. Monay, F., Gatica-Perez, D.: PLSA-based image auto-annotation: constraining the latent space. In: Proceedings of ACM International Conference on Multimedia, pp. 348–351 (2004)
10. Peng, Y.X., Lu, Z.W., Xiao, J.G.: Semantic concept annotation based on audio PLSA model. In: Proceedings of ACM International Conference on Multimedia, pp. 841–844 (2009)
11. Rainer, L., Stefan, R., Eva, H.: Multilayer pLSA for multimodal image retrieval. In: Proceedings of the ACM International Conference on Image and Video Retrieval (2009)
12. Shah-hosseini, A., Knapp, G.: Semantic image retrieval based on probabilistic latent semantic analysis. In: Proceedings of ACM International Conference on Multimedia, pp. 452–455 (2004)
13. The Caltech-101 Object Categories,
<http://www.vision.caltech.edu/~feifeili/Datasets.htm>
14. Xu, W., Liu, X., Gong, Y.H.: Document clustering based on non-negative matrix factorization. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 267–273. ACM Press, NY (2003)
15. Zhang, R.F., Zhang, Z.F.: Effect image retrieval based on hidden concept discovery in image database. *IEEE Transactions on Image Processing* 16(2), 562–572 (2007)

Genre Classification and the Invariance of MFCC Features to Key and Tempo

Tom LH. Li and Antoni B. Chan

Department of Computer Science, City University of Hong Kong, Hong Kong
lihuiali2@cityu.edu.hk, abchan@cityu.edu.hk

Abstract. Musical genre classification is a promising yet difficult task in the field of musical information retrieval. As a widely used feature in genre classification systems, MFCC is typically believed to encode timbral information, since it represents short-duration musical textures. In this paper, we investigate the invariance of MFCC to musical key and tempo, and show that MFCCs in fact encode both timbral and key information. We also show that musical genres, which should be independent of key, are in fact influenced by the fundamental keys of the instruments involved. As a result, genre classifiers based on the MFCC features will be influenced by the dominant keys of the genre, resulting in poor performance on songs in less common keys. We propose an approach to address this problem, which consists of augmenting classifier training and prediction with various key and tempo transformations of the songs. The resulting genre classifier is invariant to key, and thus more timbre-oriented, resulting in improved classification accuracy in our experiments.

1 Introduction

Musical information retrieval is a field that is growing vigorously in recent years thanks to the thriving digital music industry. As a promising yet challenging task in the field, musical genre classification has a wide-range of applications: from automatically generating playlists on an MP3 player to organizing the enormous billion-song database for major online digital music retailers. In many genre classification systems, the Mel-frequency cepstral coefficients (MFCCs) [3] have been used as a timbral descriptor [15][12][10][7]. While it is common to think of MFCCs as timbre-related features, due to the short-duration frame on which they are extracted (e.g., 20 milliseconds), it is still uncertain how the key and tempo of a song affects the MFCC features, and hence the subsequent genre classification system.

In this paper, we attempt to address the following question: are MFCCs invariant to key and tempo? In other words, is MFCC a purely timbral feature set? If the MFCCs are purely timbral features, then they should be invariant to the changes in musical keys and tempo. Otherwise, changes in the musical key and tempo of a song will affect the MFCCs, which may adversely affect the training of genre classifiers. The contributions of this paper are three-fold. First, we show that musical genres, which *should* be independent of key, are in

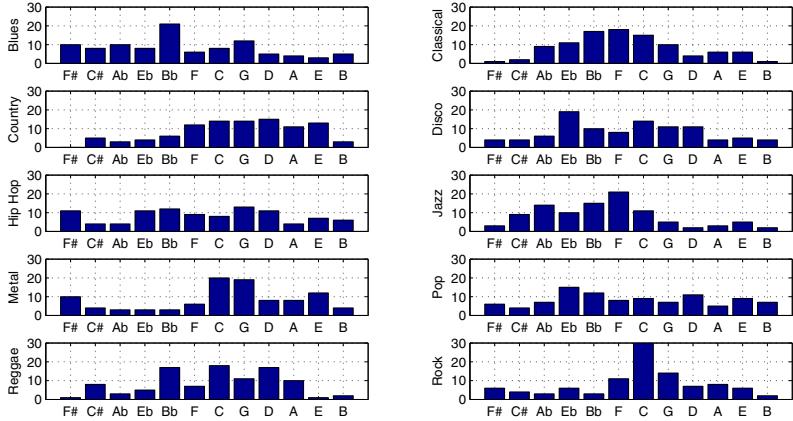


Fig. 1. Key histograms of the GTZAN dataset on the circle of fifths scale. The vertical axis is the number of songs with a certain key.

fact influenced by the fundamental keys of the instruments involved. Second, we show that MFCCs indeed encode both timbral and key information, i.e., they are not invariant to shifts in musical key. As a result, genre classifiers based on the MFCC features will be influenced by the dominant keys of the genre, resulting in poor performance on songs in less common keys. Third, we propose an approach to build key-independent genre classifiers, which consists of augmenting the classifier training and prediction phases with various key and tempo transformations of the songs. The resulting genre classifier is invariant to key, and thus more timbre-oriented, resulting in improved classification accuracy in our experiments.

The rest of this paper is organized as follows. In Section 2, we explore the distribution of musical key for different genres. In Section 3, we study the invariance of MFCC to musical key and tempo shifts. In Section 4, we propose a data-augmented genre classification scheme, based on key and tempo transformations, while in Section 5 we present experiments on genre classification using our data-augmented system.

2 Key Histograms of the GTZAN Dataset

In this section, we explore the relationship between musical genres and musical keys. We manually annotate each song in the GTZAN dataset [15] with their musical “keys”. In this section, we define the concept of “key” as the pitch of the “Do” sound of the song in the solfège scale (Do-Re-Mi scale). Such definition is different from the more common definition — the tonic sound of the scale (e.g., in minor scales the tonic sound is the La sound rather than the Do sound). Because a major scale and its relative minor scale share the identical composition of pitches, it is simpler to annotate both scales with the same label to show that

they actually have the same pitch ingredients in the songs (e.g., songs in C major and A minor are both labeled with “C”). In cases where the scale is unapparent, we annotate the key based on the most repeated pitch.

Figure II shows the key histograms for different genres in the GTZAN dataset, using our annotation criteria, with keys ordered by the circle of fifths (C is in the center). We observe that genre is indeed key-related with the distribution centered around particular keys based on the instrumentation.

- Blues: peaks at B \flat and G. B \flat is the fundamental pitch of many horn instruments. G corresponds to the Do sound for the blues scale in E, which is the fundamental key for guitar.
- Classical: distribution around F, which is in between the horn instrument fundamental B \flat and the piano fundamental C.
- Country: broad distribution around D, with keys that are easy to play on guitars (e.g. G, D, A, E, C).
- Disco: peaks at E \flat and C. Disco frequently employs Blues scale. For C Blues, the Do sound is E \flat .
- Hip Hop: distribution is not obvious. This genre typically does not have a key, as the main instruments are human voice and drums.
- Jazz: distribution is skewed towards flat keys (D \flat , A \flat , E \flat , B \flat), which are the fundamental horn pitches. The peak at F is similar to that of Classical.
- Metal: peaks at C, G, E and F \sharp . The G key correspond to E Blues. E is the pitch of the lowest string on guitar. In Metal, the lowest string is used extensively to create a massive feeling. The peak at F \sharp , corresponding to E \flat Blues, can be explained by the common practice of Metal artists to lower the tuning by one semi-tone, creating an even stronger metal feeling.
- Pop: distribution is not obvious. The peak at E \flat is the Blues-scale of the C key. The distributions of Pop and Disco are similar, due to similar instrumentation.
- Reggae: peaks at C (keyboard), D (guitar), B \flat (horns) and C \sharp (B \flat Blues).
- Rock: significant distribution around C. The distribution is related to the dominance of guitar and piano in this genre. Rock is arguably the most key-related genre in the GTZAN dataset.

In summary, there is a strong correlation between genre and key, with each genre having a unique key distribution. Such correlation most likely stems from the fundamental keys associated with the instruments used in each genre. For instance, the most common kind of clarinet is in the key of B \flat , while the alto saxophone is in E \flat . The four strings of a violin are tuned by standard to G, D, A and E. The piano has all its white keys in C major. Although it is entirely possible to play a song in any key, some keys are arguably easier to play than others, depending on the instruments used. Hence, the key characteristics of instruments could unexpectedly associate musical keys to specific genres.

3 Are MFCCs Invariant to Key and Tempo?

In this section we study the invariance of MFCCs to shifts in musical key and tempo.

3.1 Mel-Frequency Cepstral Coefficients

The mel-frequency cepstral coefficients (MFCC) [3] are a widely adopted audio feature set for various audio processing tasks such as speech recognition [13], environmental sound recognition [9], and music genre classification [15, 2, 7, 11]. [11] investigated the MFCC features on various time scales and with different modeling techniques, such as autoregressive models. [15, 8] compared the MFCCs to the short-time Fourier transform (STFT), beat histogram and pitch histogram feature sets, concluding that MFCCs give best performance as an independent feature set.

Given a frame of audio, the computation of MFCC involve the following steps that mimic the low-level processing in the human auditory system [3]: 1) transformation of the audio frame to the frequency domain using the STFT; 2) mapping the frequency bins to the mel-scale, using triangular overlapping windows; 3) taking the logs of the mel-band responses; 4) applying a discrete cosine transform (DCT) to the mel-bands. In this paper, the MFCCs are extracted with the CATBox toolbox [4], using 40 mel-bands and 13 DCT coefficients. The frame size is 18 milliseconds, taken every 9 milliseconds.

3.2 Key and Tempo Transformations

To examine the changes of MFCC values to shifts in keys and tempos, we apply key shifting and tempo shifting musical transforms to each song in the GTZAN dataset. These transformations consist of sharpening/flattening the song up to 6 semitones, and changing the tempo 5% and 10% faster/slower. The transformations are performed with the WSOLA algorithm [16], which is implemented in the open-source audio editor Audacity [1]. The musical transforms are analogous to affine transforms of images, which deform an image without changing the general shape (e.g. rotating and skewing the number 1). Augmenting the dataset with affine transforms is a common technique in digit recognition tasks [14], where the enlarged training set improves classification accuracy by encouraging invariance to these deformations.

There are doubts that transforming a song to approximate the key-shifted and tempo-shifted version of the songs might not be appropriate, since such transforms might also contaminate the timbral characteristics of the songs. We argue that such an effect is minor for the following three reasons: 1) qualitatively speaking, the transformed songs sound perceptually very similar to the original song recorded in different key and tempo, with critical information for genre classification, such as instruments, musical patterns and rhythm characteristics, still preserved; 2) considering that musical instruments have different timbre in different registers, we limit the key shifts to the range of half an octave (from $\flat 6$ to $\sharp 6$); 3) we compared the MFCC values extracted from MIDI songs and their perfect key-transposed versions, and observed that the MFCC values vary in similar ways as in the key-transformed songs.

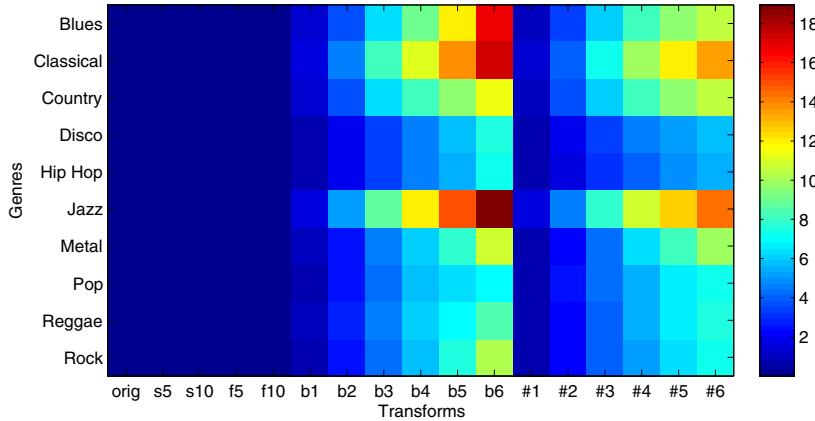


Fig. 2. MFCC KL-divergence: the horizontal axis represents the key and tempo transforms, from left to right, original, 5% slower, 10% slower, 5% faster, 10% faster, key transform b_1 to b_6 and $\#_1$ to $\#_6$. The color represents the average KL divergence between corresponding frames in the original and transformed songs.

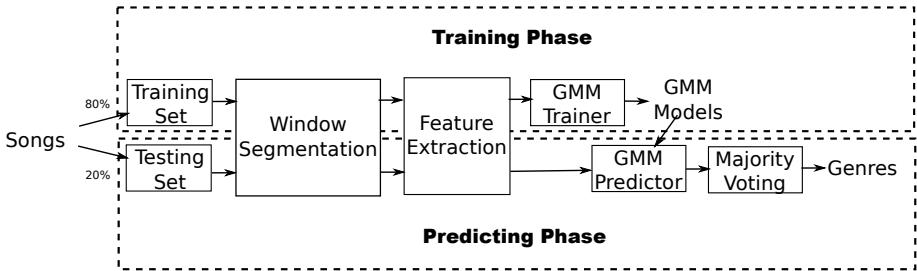
3.3 Comparison of MFCCs under Key and Tempo Transforms

For genre classification, MFCCs are often aggregated over a long-duration window using statistical methods [15][2]. Motivated by this fact, we compare the original songs and their transformed versions by computing the Kullback-Leibler (KL) divergence [5] between corresponding windowed excerpts (3.5 seconds). Assuming that the MFCCs in a window follow a Gaussian distribution (e.g., as in [15]), the calculation of KL divergence between two windows is given by:

$$D_{KL}(\mathbf{N}_0 \parallel \mathbf{N}_1) = \frac{1}{2} \left(\log \frac{|\Sigma_1|}{|\Sigma_0|} + \text{tr}(\Sigma_1^{-1} \Sigma_0) + (\mu_1 - \mu_0)^\top \Sigma_1^{-1} (\mu_1 - \mu_0) - d \right) \quad (1)$$

where (μ_0, Σ_0) and (μ_1, Σ_1) are the mean and covariance for the two Gaussian distributions, and d is the dimension.

Figure 2 shows the KL divergence between different musical transforms of the same songs, averaged over each genre. From the figure, we see that key transforms affect the MFCC distribution, with larger key shifts affecting the distribution more. Interestingly, MFCCs for some genres are more sensitive to the changes in key, such as blues, jazz and metal. This can be explained by the fact that these genres have instruments with richer harmonic structure, and therefore the MFCCs change more since they model timbre. On the other hand, tempo transforms do not have a great effect on the distribution of MFCC values. This is because transforming a song in time does not change the frequency characteristics, but only the number of MFCC frames. Compressing a song subsamples the MFCC frame set, while stretching it adds new MFCC frames by interpolation. In both cases, the distribution of the MFCCs over the window remains about the same.

**Fig. 3.** System architecture

In the previous, we showed that genres have dominant keys, due to the instrumentation of the genre. On the other hand, in this section, we have shown that MFCCs, which are common features for genre classification, are not invariant to key transformations. This brings forward an interesting dilemma. Because genre is key dependent and MFCCs are not key invariant, then a classifier based on MFCCs may overfit to the dominant keys of the genre. The resulting classifier will then have poor accuracy on songs in the less common keys. In the next section, we look at learning a key-invariant genre classifier, by augmenting the classifier with different musical transforms.

4 Genre Classification with Musical Transforms

In this paper, we adopt the genre classification system of [15][2][11]. Figure 3 shows the architecture of the system, which contains four steps. First, the input song is split into non-overlapping windows of equal length (as in [2], we use window length of 3.5 seconds). These windows then go through a feature extraction process, producing feature vectors which are compact representations of those windows. In particular, MFCCs are first extracted from the audio signal, and the mean and standard deviation of the MFCCs over the window are calculated as the feature vector. In the third step, the feature vector is fed to a Gaussian mixture model (GMM) classifier. The parameters of the GMM classifier are learned from the training set using the EM algorithm [6], which iteratively estimates the parameters by maximizing the likelihood of the training set. One GMM is learned for each genre. Given a feature vector extracted from a window, the GMM with the largest likelihood is selected as the genre label for the window. The labels for all the windows in a song are then aggregated with a majority voting process to produce a genre label for the song.

We can modify the genre classification system in two ways to make it invariant to musical transforms. First, in the training phase, we can expand the training set by adding transformed versions of the training songs, hence generating more examples for learning the genre classifier. Second, in the prediction phase, we can augment the classifier by processing the test song along with its transformed versions. The final label for the test song is the majority vote over all windows of all versions of the songs. The data augmentation step can be seen as adding

a sample diffusion layer before either the training or the predicting phase of the system.

5 Experiments

In this section we present our experimental results on genre classification in the context of key and tempo augmentation.

5.1 Dataset and Experimental Setup

In our experiments, we use the GTZAN dataset [15], which contains 1000 song clips of 30 seconds each, with a sampling rate of 22050 Hz at 16 bits. There are 10 musical genres, each with 100 songs: Blues, Classical, Country, Disco, Hip hop, Jazz, Metal, Pop, Reggae, and Rock. We augment the original GTZAN dataset (denoted as the “Orig” dataset) using different combinations of musical transforms. The “Tempo” dataset contains the Orig dataset and its tempo variants, 5% and 10% faster/slower. The “Key” dataset contains the Orig dataset and its key variants from $\flat 6$ to $\sharp 6$. The “Tempokey” dataset is the union of the Tempo and Key datasets. We also augment our dataset with key transforms that are based on the circle of fifths. The “Fifth1” dataset contains the Orig dataset and its key variants with one step on the circle of fifths, i.e. $\flat 5$ and $\sharp 5$, while the “Fifth2” dataset contains variants with one more step, i.e. $\flat 2$ and $\sharp 2$. The circle of fifths augmented datasets are strict subsets of the Key dataset.

We carried out three different sets of experiments in combination with the 6 augmentations listed above. In the first experiment, denoted as AugTrain, the classifiers are trained using the augmented dataset, while genre prediction is performed using only the original songs. In the second experiment, denoted as AugPredict, the classifiers are trained only on the original dataset, while prediction is performed by pooling over the augmented song data. In the final experiment, denoted as AugBoth, both the classifier training and prediction use the augmented song data. Genre classification is evaluated using five random splits of the dataset, with 80% of the songs (and its variants) used for training, and the remaining 20% used for testing. The experiments are carried out on a range of parameters. We use MFCC lengths from 1 to 13 (i.e., the number of DCT coefficients), and vary the number of components in the GMM (K) from 1 to 20. We also assume diagonal covariance matrices in the GMM.

5.2 Experimental Results

We first examine the effects of the system parameters, such as the size of the GMM and the length of the MFCCs. Figure 4a shows the classification accuracy, averaged over all the data augmentations and MFCC lengths, while varying the number of components in the GMM. In general, the classification accuracy increases with K , and there does not seem to be an over-fitting problem for large K , such as 20. Figure 4b shows the accuracy, averaged over all data augmentations and GMMs, while varying the length of the MFCCs. Similarly, the

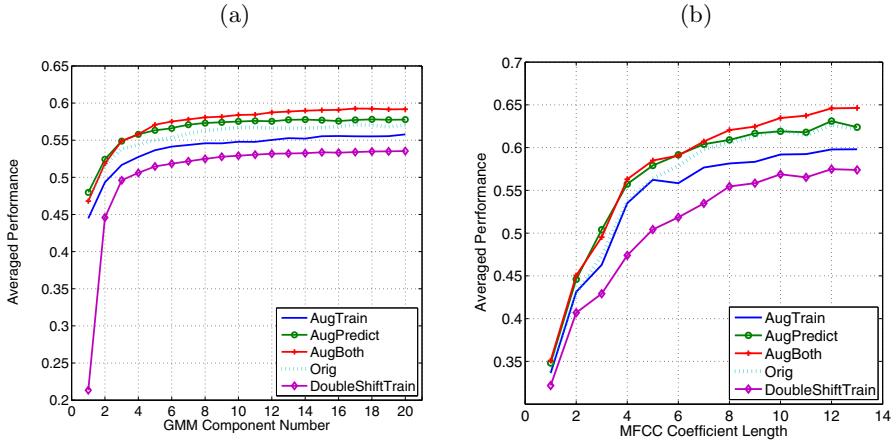


Fig. 4. (a) Averaged accuracy for all datasets and MFCC lengths, while varying the number of GMM components (K); (b) Averaged accuracy for all datasets and GMM components, while varying the MFCC length

accuracy improves as more MFCCs are added. In fact, despite their sensitivity to noise, these high-order coefficients provide useful details for genre classification. As a comparison, [15] limited their system to the first 5 MFCC coefficients and GMMs with $K=5$, and achieved 61% classification accuracy when using MFCCs with three other types of features. In contrast, our system scores 66.3% on the Orig dataset when using 13 MFCC features.

Next, we look at the effect of signal degradation when using the music transformation. In particular, we add noise to the Orig training set by applying a “double-shift” to each training song. This consists of first shifting the key of the song, and then shifting it back to the original scale. The result is a training song with noise added due to the musical transformation. The double-shifted training set is used to train the genre classifier, which then predicts genres on the Orig test data. This result is denoted as DoubleShiftTrain in Figure 4. In particular, using the noisy training data degrades the accuracy, when compared to the Orig performance (e.g., the accuracy drops 5% to 53.5% for $K=20$). However, in spite of this added noise to the training set, the system is still able to do genre classification, albeit with reduced accuracy.

Finally, we look at the effect of using the proposed data-augmented classifiers. From Figure 4, we observe that the AugTrain classifier gives constantly better performance than the DoubleShiftTrain classifier, while its performance is still lower than that of the Orig dataset. This suggests that using augmented training data improves the accuracy, at least compared to the unaugmented classifier using similar noisy training data. This improvement, however, is not enough to overcome the transformation noise. On the other hand, using data-augmented prediction (AugPredict) gives constantly better performance than the Orig dataset. Finally, using both data-augmented classification and prediction (AugBoth) achieves the best accuracy, dominating both AugPredict and

Table 1. Genre classification accuracy for different data-augmentation schemes and transformed datasets, for K=20 and MFCC length 13

	Tempo	Key	Tempokey	Fifth1	Fifth2	Average
Orig	—	—	—	—	—	64.5%
DoubleShiftTrain	—	—	—	—	—	61.9%
AugTrain	65.1%	62.0%	64.5%	60.5%	62.8%	63.0%
AugPredict	66.2%	63.6%	66.4%	61.0%	63.7%	64.2%
AugBoth	66.6%	67.8%	68.9%	67.5%	67.3%	67.6%

Table 2. AugBoth Classification Rates for different genres, with K = 20 and MFCC length 13

	blues	classical	country	disco	hip-hop	jazz	metal	pop	reggae	rock	average
Orig	59	92	62	41	64	86	77	58	61	45	64.5
Tempo	64	97	62	46	66	85	75	64	68	39	66.6
Key	62	99	67	55	65	90	83	64	60	33	67.8
Tempokey	63	98	67	55	65	91	87	61	63	39	68.9
Fifth1	61	98	67	52	63	88	83	63	62	38	67.5
Fifth2	64	94	63	58	63	90	79	64	66	32	67.3

Orig. Table 1 shows the average classification accuracy using different transformed datasets and data-augmentation schemes for K=20 and MFCC length 13. The best performance achieved for all experiments is 69.3%, using the AugBoth classifier with the Key transformations, K=18 and MFCC length 13.

Table 2 shows the classification accuracy for different genres using the Aug-Both classifier. Comparing the genres, Classical has the highest accuracy, scoring over 90% on all datasets, followed by Jazz and Metal. In contrast, Disco and Rock are the two worst performing genres. In general, the augmentation of the dataset improves the genre classification. The only exception is the Rock genre, where augmentation always lowers the classification accuracy. Looking at the confusion matrix for AugBoth, we found that more instances of Rock are misclassified as Metal. On the other hand, Disco performs significantly better because less instances are misclassified as Blues, Pop and Rock.

5.3 Discussion

From these experimental results we have three conclusions. First, the MFCC feature set is largely a timbral feature set. From the confusion matrices we found that confusable genres have similar instrumentation. Additionally, genres with distinct instrumentation stand out from others easily, e.g., Classical uses orchestral instruments, while Metal has high frequency distorted guitar.

Second, in addition to timbral information, MFCCs also encodes key information, which eventually affects the genre classification accuracy. We observed that the key and tempo augmented classifiers have a significant change in performance over the baseline. Rock and Metal both use guitars and drums as the

main instruments, but they have very different key distributions as shown in Figure 11. The confusion between Rock and Metal after key augmentation suggest that the classification of Rock music is partly due to musical keys. If we blur the lines between keys for these two genres, we are likely to lose such information, leading to a degradation of classification performance.

Third, making the genre classifier tempo- and key-invariant, via data augmentation, generally improves the classification accuracy. The accuracies of the AugTrain, AugPredict and AugBoth classifiers are significantly better than the noise-added DoubleShiftTrain baseline. Despite the noise from the imperfect musical transforms, the accuracy of the AugPredict and AugBoth classifiers are constantly better than the Orig baseline. These results suggest a method for boosting overall genre classification performance, by artificially generating transformed songs to augment the classifier training and prediction phases, thus strengthening the timbre-orientation of the classifier. However, some genres (e.g. Rock) will suffer from such augmentation since the recognition of that genre is partly due to musical keys.

While the concept of “musical genre” is perceptual and largely based on timbre information, there is still a strong correlation between genre and key, due to instrumentation, which should also be considered. Future work will look at combining timbral and key information, using appropriate machine learning models, to push the performance further. In addition, reducing the noise introduced by the musical transform will also likely improve the classification accuracy.

6 Conclusion

MFCCs are widely used audio features in music information retrieval. Extracted over a short-duration frame, MFCCs are typically perceived as a timbral descriptor. In this paper, we have shown that the MFCCs are not invariant to changes in key, and hence they encode both timbral and key information. On the other hand, we found that musical genres, which should be independent of key, are in fact influenced by the fundamental keys of the instruments involved. As a result, genre classifiers based on the MFCC features will be influenced by the dominant keys of the genre, resulting in poor performance on songs in less common keys. We suggested an approach to address this problem, which consists of data-augmentation during the classifier training and prediction phases, with key and pitch transformations of the song. The resulting genre classifier is invariant to key, and thus more timbre-oriented, resulting in improved classification accuracy in our experiments.

References

1. Audacity, the free, cross-platform sound editor, <http://audacity.sourceforge.net/>
2. Bergstra, J., Casagrande, N., Erhan, D., Eck, D., Kégl, B.: Aggregate features and Adaboost for music classification. *Machine Learning* 65(2), 473–484 (2006)

3. Bridle, J.S., Brown, M.D.: An experimental automatic word recognition system. JSRU Report, 1003 (1974)
4. Computer audition toolbox, <http://cosmal.ucsd.edu/cal/projects/catbox/catbox.htm>
5. Cover, T.M., Thomas, J.A.: Elements of information theory. John Wiley and Sons, Chichester (2006)
6. Dempster, A.P., Laird, N.M., Rubin, D.B., et al.: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1), 1–38 (1977)
7. Ellis, D.: Classifying music audio with timbral and chroma features. In: Int. Symp. on Music Information Retrieval (ISMIR), pp. 339–340 (2007)
8. Li, T., Tzanetakis, G.: Factors in automatic musical genre classification of audio signals. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 143–146 (2003)
9. Lu, L., Zhang, H.J., Li, S.Z.: Content-based audio classification and segmentation by using support vector machines. *Multimedia Systems* 8(6), 482–492 (2003)
10. Mandel, M., Ellis, D.: Song-level features and support vector machines for music classification. In: Proc. ISMIR, pp. 594–599. Citeseer (2005)
11. Meng, A., Ahrendt, P., Larsen, J.: Improving music genre classification by short time feature integration. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005), vol. 5 (2005)
12. Pachet, F., Aucouturier, J.J.: Improving timbre similarity: How high is the sky?. *Journal of negative results in speech and audio sciences* 1(1) (2004)
13. Pearce, D., Hirsch, H.G.: The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Sixth International Conference on Spoken Language Processing. Citeseer (2000)
14. Simard, P.Y., Steinkraus, D., Platt, J.: Best practices for convolutional neural networks applied to visual document analysis. In: International Conference on Document Analysis and Recognition (ICDAR), pp. 958–962. IEEE Computer Society, Los Alamitos (2003)
15. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. *IEEE Transactions on speech and audio processing* 10(5), 293–302 (2002)
16. Verhelst, W., Roelands, M.: An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech. In: IEEE International Conference on Acoustic Speech and Signal Processing, vol. 2, Institute of Electrical Engineers Inc., (IEE) (1993)

Combination of Local and Global Features for Near-Duplicate Detection

Yue Wang, ZuJun Hou, Kariant Leman, Nam Trung Pham, TeckWee Chua,
and Richard Chang

Institute for Infocomm Research

A*Star (Agency for Science, Technology and Research), 1 Fusionopolis Way, Singapore
`{ywang, zhou, kariant, ntpham, tewchua, rpchang}@i2r.a-star.edu.sg`

Abstract. This paper presents a new method to combine local and global features for near-duplicate images detection. It mainly consists of three steps. Firstly, the keypoints of images are extracted and preliminarily matched. Secondly, the matched keypoints are voted for estimation of affine transform to reduce false matching keypoints. Finally, to further confirm the matching, the Local Binary Pattern (LBP) and color histograms of areas formed by matched keypoints in two images are compared. This method has the advantage for handling the case when there are only a few matched keypoints. The proposed algorithm has been tested on Columbia dataset and compared quantitatively with the RANdom SAmple Consensus (RANSAC) and the Scale-Rotation Invariant Pattern Entropy (SR-PE) methods. The results turn out that the proposed method compares favorably against the state-of-the-arts.

Keywords: Near-duplicate detection, image matching, LBP histogram, color histogram, keypoints, affine invariant feature.

1 Introduction

Near-Duplicate (ND) image detection aims to evaluate a pair of images which one is close or partially close to the other. It has been widely applied to image and video content analysis, such as location recognition, redundant contents detection, image spam detection and illegal copy of images and videos detection. The existing techniques can roughly be classified into two categories:

- 1) Local methods (keypoint-based methods) [1-8]: which detect local keypoints in two images and measures their similarity by counting the number of correct correspondences between two sets of keypoints. Most widely used methods for keypoint detection and matching include SIFT, SURF, and various extensions.
- 2) Global methods (appearance-based methods) [9-12]: which employ global features from whole images, such as color moments, color histogram, Gabor feature, local binary pattern feature, or image edges.

Usually local methods can deal with the illumination variation and geometric transformation but at the expense of computational efficiency. Comparatively, global methods are very efficient in coding and ideal for finding identical copies, but tend to

be sensitive to the variation of lighting and viewpoint, or occlusions. Mikolajczyk and Schmid have conducted a comparison of several local descriptors [14]. For a brief account of the latest development on ND image detection, interested readers can refer to [9] or [8]. This paper will present a method to combine the advantages of appearance-based method and keypoint-based method for affine ND image detection. The proposed method firstly conducts the local keypoint detection and matching, and then estimates affine transform based on a difference of affine invariant ratios of areas, finally performs Local Binary Pattern (LBP) and color histogram matching to further confirm the content within the region of interest as located by the local matching. The method is advantageous in handling the case when there are only a few matched keypoints.

The remaining structure of this paper is arranged as follows. Our method is detailed in Section 2. Experimental results and discussion are presented in Section 3. Finally, the paper is concluded in Section 4.

2 Methods

The basic idea of our method is to combine local and global feature for ND image detection. A flowchart is illustrated in Fig. 1. The method starts from a local method for preliminary detection and matching. After that, global information is gradually added to further enhance the decision. It is based on filtering with affine invariant feature, followed by LBP and color histogram confirmation. We assume two ND images are related through an affine transformation with certain illumination variation.

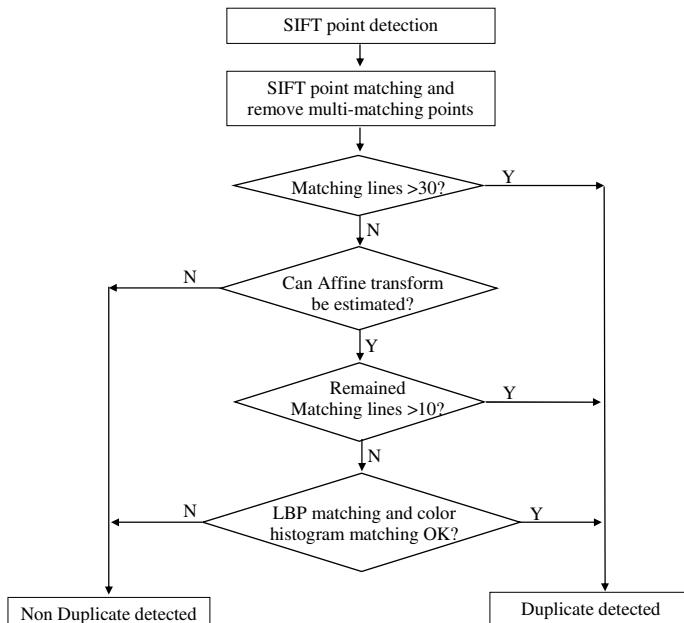


Fig. 1. Flowchart for complete algorithm

2.1 Keypoint Detection and Matching

In computer vision, there have been many methods to detect and match salient points. This study adopts the SIFT, which computes a histogram of local oriented gradients and stores the bins as feature vector for matching. Normally there are many matching lines for two duplicate images, especially for the case that two images are only slightly changes in rotation and scale, there are many matching lines can be observed, see Fig. 2(a) for an example. In this case, a simple threshold for the number of matching lines is able to achieve a good performance. However, when two images are related with significant change in rotation, scale or illumination, the number of matching lines could be as small as few (e.g. Fig. 2(b)). On the other hand, there could be cases where point pairs are falsely matched. Thus, the local information could be insufficient for ND image detection. In the following we present how these problems could be handled through an analysis of more global feature.

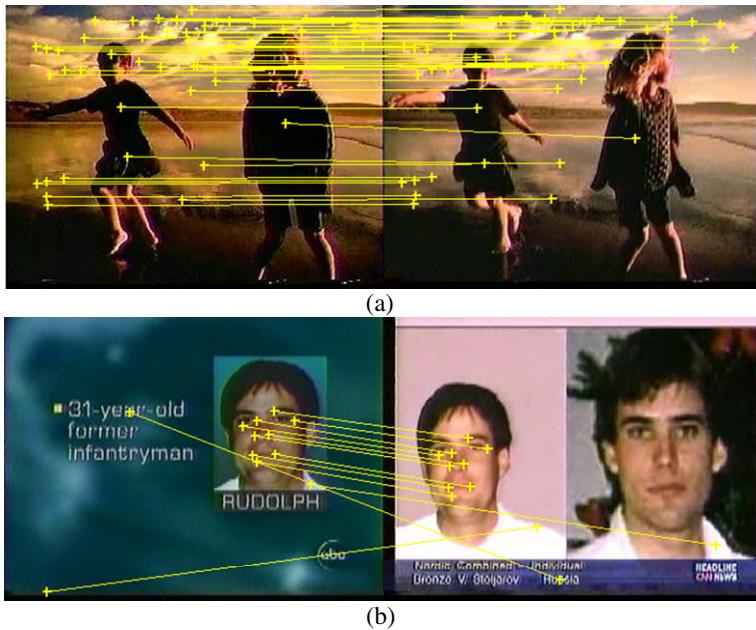


Fig. 2. Keypoints matching. (a) Many matching lines; (b) Fewer matching lines.

2.2 Matching Lines Filtering Based on Affine Invariant Feature

In general, an affine transformation consists of a linear transformation (rotation, scaling or shear) followed by a translation:

$$\mathbf{T}(\mathbf{x}) : \mathbf{x} \mapsto \mathbf{Ax} + \mathbf{b}. \quad (1)$$

There are some properties for two images if they are related with an affine transformation in Euclidean space, which include the preservation of the collinearity relation between points as well as the ratio of distances for distinct collinear points. Parallel lines will remain parallel. Most interestingly, the ratio of the area will be a

constant. When the determinant of \mathbf{A} is 1 or -1, the area will be preserved as well. This property has been utilized in [15] for image matching, where the ratio of areas are formed by three matched pairs.

In Fig. 3, two sets of points $S = \{A, B, C, D, E, F\}$ and $S' = \{A', B', C', D', E', F'\}$ are related with an affine transformation: $T(S) = S'$. Based on the property of affine transformation, we have

$$\frac{\text{Area}_{ABC}}{\text{Area}_{DEF}} = \frac{\text{Area}_{A'B'C'}}{\text{Area}_{D'E'F'}} = \alpha \quad (2)$$

where α is a constant. Eq. (1) can be written as

$$R_{diff} = R - R' \quad (3)$$

where

$$R = \frac{\text{Area}_{ABC}}{\text{Area}_{DEF}}, \quad R' = \frac{\text{Area}_{A'B'C'}}{\text{Area}_{D'E'F'}} \quad (4)$$

For the case of affine transformation, Eq. (3) should be

$$R_{diff} = 0 \quad (5)$$

Therefore α can be ignored in Eq. (3).

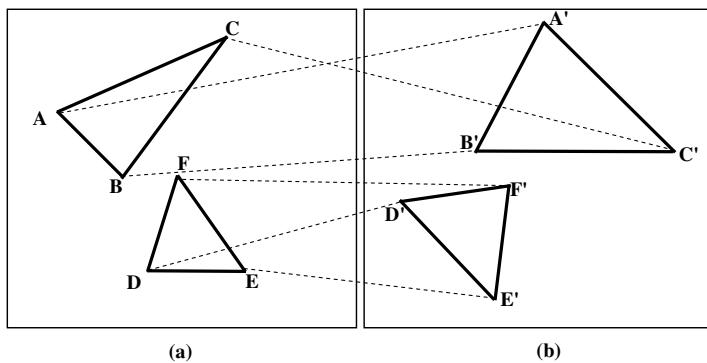


Fig. 3. Affine transformed triangles

As aforementioned, the affine transformation preserves the ratio α , thus, those triangles under the same affine transformation will lead to zero ratio difference in Eq. (3), indicating the similarity among the keypoints associated with these triangles. On the other hand, those triangles under different affine transformation will yield different ratios and lead to non-zero ratio difference in Eq. (3), which suggests that these keypoints are less correlated and these matching pairs are more likely a false matching. If we create a voting space with some number of categories (bins) and count the occurrence frequency of each category, we will arrive at a histogram as shown in Fig. 4. In this graphical display, areas under the same affine transformation

will vote for the bin with R_{diff} equals to zero and the votes of areas under different affine transformation will scatter among different bins, appearing like background/noise in the histogram, thus we can inference the degree of ND from the analysis of this histogram. If a peak is identified with significantly large value in bin of zero, the confidence on the near-duplicate between two images would be high. Otherwise, the two images will be less likely near-duplicate. If we regard the peak as “signal” and others as “noise”, then a simple measure of the significance is the “signal to noise” ratio. For notational convenience, let β_i denote the i -th bin and $f(\beta_i)$ the corresponding occurrence frequency. The strength of signal is represented by the frequency of the bin $f_s = f(\beta_z)$, where β_z represents the zero of Eq. (3), and the strength of noise is characterized by f_n , the average of frequencies excluding f_s . Then the degree of near-duplicate is evaluated by

$$\gamma = \frac{f_s}{f_n}. \quad (6)$$

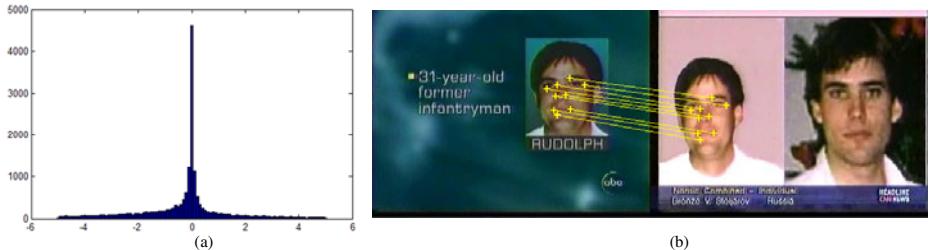


Fig. 4. Affine transform estimation. (a) Histogram of affine transformation for Fig. 2(b), where the horizontal axis is the value of R_{diff} and the vertical axis is the occurrence frequency. (b) Matching lines remained after affine transformation estimation for Fig. 2(b).

In our implementation, the number of bins in the histogram is set to 100 to handle the range of $(-5, 5)$ for R_{diff} . With this design, it is able to check either all or parts of the matching in two images. It was experimentally determined that detection of a true peak over the noise requires the peak at β_z to be at least 10 times greater than the noise level. In our application, the histogram is created with exhaust sampling of the matched keypoints. Those keypoints which contribute more than 50% of their total contribution to bin β_z are remained to estimate the affine transformation by minimum mean square error.

2.3 Confirmative Matching Using LBP and Color Histogram

As aforementioned, the number of matched lines could be very small so that the confidence of ND detection based on the number of matched lines would be low. To address this issue, LBP [17] and color histogram matching are employed for further confirmation. There are various methods for appearance and color similarity measurement. Here LBP and color histogram with Bhattacharyya coefficient [18] is

employed. LBP is originally designed for gray images as an illumination invariant texture descriptor, it has good performance in various applications such as texture classification and segmentation, face recognition and image retrieval. LBP is for gray image and intends to extract the local gradient changes, to be fully using the color information, the color histogram is generated and compare as well for enhancement of matching.

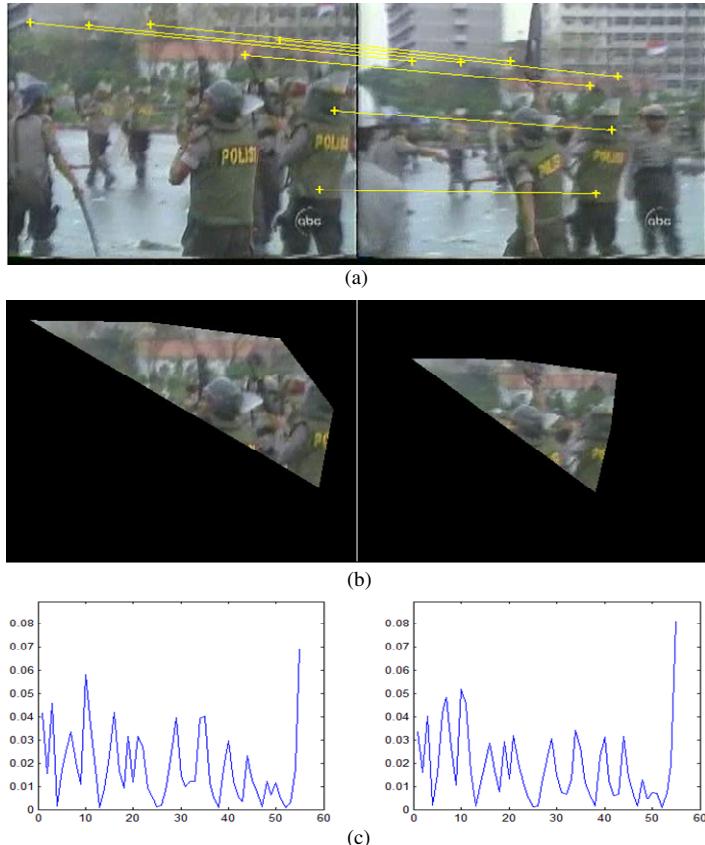


Fig. 5. LBP and color histogram matching. (a) Original images with matching lines. (b) The areas for LBP and color histogram matching. (c) Normalized LBP histograms for two regions to be compared. The measurement of S_{LBP} and S_{color} for two showed regions are 0.9927 and 0.8687 respectively.

Referring to Fig. 5, the regions of interest (ROI) formed by matched keypoints are extracted from two images respectively. First LBP is applied to the ROIs. We use uniform pattern in a neighborhood of 8 sampling points. Two LBP features' histograms can be achieved, each has 59 elements. To reduce the homogenous effect, the first 4 elements of LBP histogram are removed, leading to two normalized LBP histograms with 55 elements each, (h_i and l_i , $i=1$ to 55). The similarity of two

normalized LBP feature histograms based on Bhattacharyya coefficient can be computed by

$$S_{LBP} = \sum_{i=1}^{55} \sqrt{h_i \cdot l_i} \quad (7)$$

The value returned is from 0 to 1. A threshold of 0.93 is utilized in this study, above which will be identified as similar in LBP feature.

In addition, the color histogram is generated with 16 bins for each color channel. Then, a matrix of size 16x16x16 is constructed and reshaped as a one dimension vector with 4096 elements. For comparing two histograms, (c_i and v_i , $i=1$ to 4096), the color similarity measurement based on Bhattacharyya coefficient is as follows,

$$S_{color} = \sum_{i=1}^{4096} \sqrt{c_i \cdot v_i} \quad (8)$$

The value returned is from 0 to 1. A threshold of 0.36 is utilized in this study, above which will be identified as similar in color.

4 Experimental Results and Discussion

The Columbia dataset, which is selected from TRECVID 2003 video corpus [16], is utilized to validate the effectiveness of the proposed method. It consists of 600 key frames with 179700 candidate pairs. In this dataset, 150 pairs are selected as ND pairs. Additional 60 pairs have been chosen in [8]. However, after careful identification, we add another 5 pairs as ND. The new groundtruth which has total 215 ND pairs can be viewed at http://www1.i2r.a-star.edu.sg/~ywang/demo/columbia_groundtruth_215.

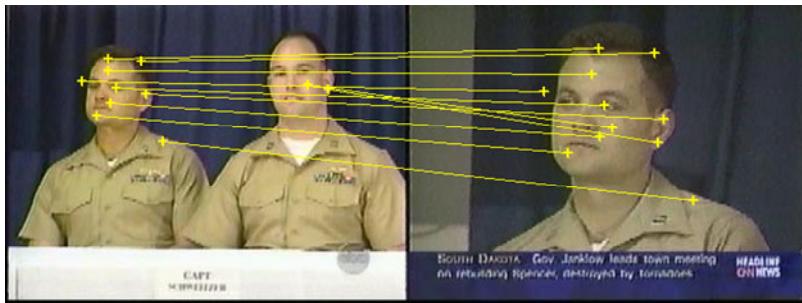
In this experiment, the ND detection is achieved by performing exhaustive search throughout the dataset. An example of correct ND detection is shown in Fig. 6. Note that the scale variation is large between the two images. It can be observed that there are some false matching lines in Fig. 6(a), but they can be successfully filtered out with the proposed method for affine transform estimation.

We compare the results of our method to RANSAC to verify the effectiveness of the proposed method for affine transform estimation. The Matlab codes for RANSAC is downloaded from [19] and the default settings are used. In addition, we also compare our method with SR-PE, which has been reported to have better performance over other methods [8]. We use the Recall, Precision and F-measure [8] to quantitatively evaluate the performance of ND detection,

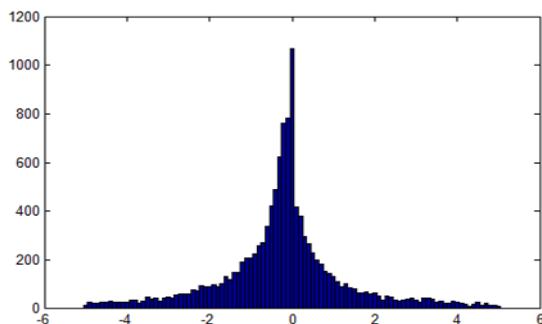
$$\text{Recall} = \frac{\text{Number of ND pairs correctly detected}}{\text{Total Number of ND pairs}} \quad (9)$$

$$\text{Precision} = \frac{\text{Number of ND pairs correctly detected}}{\text{Number of detected ND pairs}} \quad (10)$$

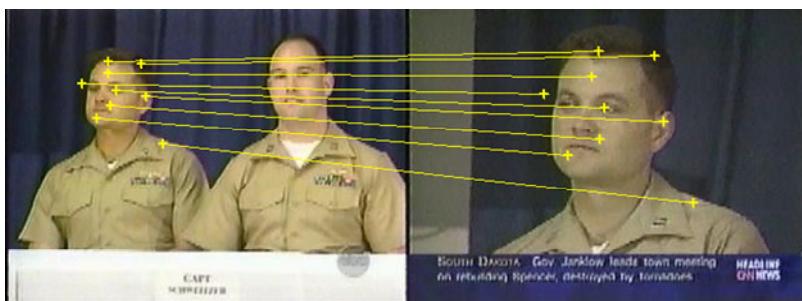
$$\text{F - measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (11)$$



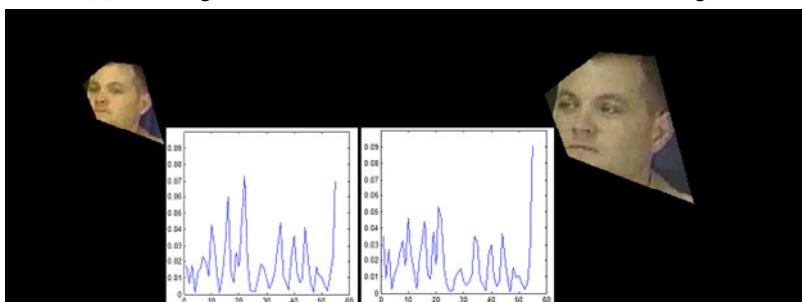
(a) Keypoints and matching before affine transformation filtering.



(b) Bin space for affine transform estimation.



(c) Matching lines remained after affine transformation filtering.

(d) Regions and LBP histogram for LBP and color matching. $S_{LBP} = 0.9888$ and $S_{color} = 0.4207$.**Fig. 6.** One example on correct matching

Recall measures the accuracy of returning ground-truth ND pairs, while Precision assesses the ability of excluding false positives. F-measure calculates the fitness of ground-truth and detected ND pairs by jointly considering recall and precision [8].

The effectiveness of the color histogram matching is firstly evaluated. Table 1 shows the results of RANSAC and the proposed method without LBP and color matching, where the number of correct ND detected is similar. The proposed method for affine transform estimation is much better than RANSAC in terms of lower false detection rate. Table 2 shows the results for both methods where LBP and color matching were applied. From there it can be observed that, the false alarms have been remarkably reduced, it shows that the proposed LBP and color matching method is effective. Table 3 shows the performance comparison of our method with SR-PE on Columbia dataset. It can be observed that although SR-PE is able to achieve a perfect precision, our method is able to get the better recall and F-measure.

Table 1. Performance of the proposed method vs RANSAC (both without LBP and color matching)

	RANSAC	Proposed method
Pairs detected	230	195
Correctly detected	178	179
Falsely detected	52	16
Precision	0.7739	0.9179
Recall	0.8279	0.8326
F-measure	0.8	0.8732

Table 2. Performance of the proposed method vs RANSAC (both with LBP and color matching)

	RANSAC	Proposed method
Pairs detected	181	181
Correctly detected	174	178
Falsely detected	7	3
Precision	0.9613	0.9834
Recall	0.8093	0.8279
F-measure	0.8788	0.8990

Table 3. Performance of the proposed method vs SR-PE on Columbia dataset

	Proposed method	SR-PE
Pairs detected	181	173
Correctly detected	178	173
False detected	3	0
Precision	0.9834	1
Recall	0.8279	0.8047
F-measure	0.8990	0.8918

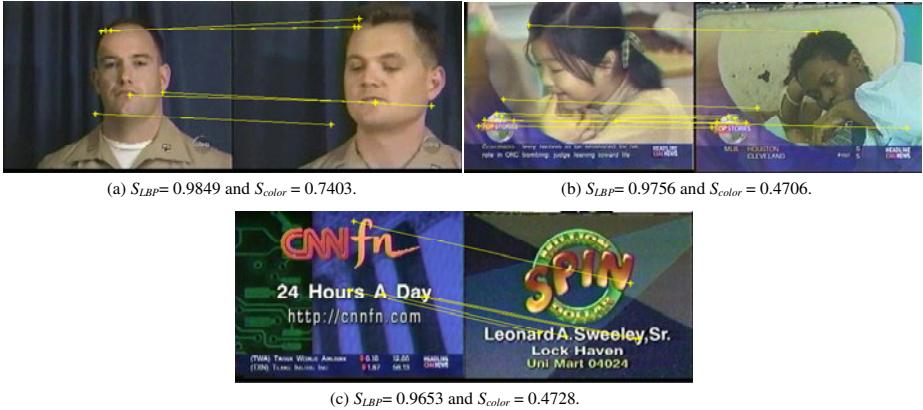


Fig. 7. Three false ND detections using the proposed method

However, there are three false alarms in our method, as showed in Fig. 7. In Fig. 7(a), the matched keypoints are well located on good corresponding places in the two images. It is worthwhile to note that these two images are not the ND in the ground truth of Columbia dataset, despite that they are similar in background and color, but with different people and scale. Fig. 7(b) is mainly affected by the left-bottom logo, while the false matching in Fig. 7(c) is mostly due to few matching lines and similar color tone in the sampled areas.

5 Conclusion

This paper presents our works for ND detection. The key component of our method is a combination of local and global features, where global information is utilized to reduce the false matching from local features, and to further enhance the matching when the number of matching lines is limited. The qualitative comparison with SR-PE shows that the proposed method is on par with the state-of-the-art. Our future work includes exploring the possibility of incorporating other methods for local matching and more specific detail to enhance the global matching in order to increase the recall and to reduce the false alarm case as showed in Fig. 7.(b)(c).

References

1. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust Wide Baseline Stereo from Maximally Stable Regions. In: British Machine Vision Conference, pp. 384–396 (2002)
2. Mikolajczyk, K., Schmid, C.: Scale and Affine Invariant Interest Point Detectors. International Journal of Computer Vision 60, 63–86 (2004)
3. Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
4. Ke, Y., Sukthankar, R.: PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In: CVPR, vol. 2, pp. 506–513 (2004)

5. Ke, Y., Sukthankar, R., Huston, L.: Efficient Near-Duplicate Detection and Sub-Image Retrieval. In: ACM Multimedia, pp. 869–876 (2004)
6. Zhao, W.L., Ngo, C.W., Tan, H.K., Wu, X.: Near-Duplicate Keyframe Identification with Interest Point Matching and Pattern Learning. IEEE Trans. on Multimedia 9(5), 1037–1048 (2007)
7. Ngo, C.W., Zhao, W.L., Jiang, Y.G.: Fast Tracking of Near-Duplicate Keyframes in Broadcast Domain with Transitivity Propagation. In: ACM Multimedia, pp. 845–854 (2006)
8. Zhao, W.L., Ngo, C.W.: Scale-Rotation Invariant Pattern Entropy for Keypoint-based Near-Duplicate Detection. IEEE Trans. on Image Processing 18(2), 412–423 (2009)
9. Zhu, J., Hoi, S.C.H., Lyu, M.R., Yan, S.: Near-Duplicate Keyframe Retrieval by Nonrigid Image Matching. In: MM 2008, pp. 41–50 (2008)
10. Qamra, A., Meng, Y., Chang, E.Y.: Enhanced Perceptual Distance Functions and Indexing for Image Replica Recognition. PAMI 27(3), 379–391 (2005)
11. Zhang, D.Q., Chang, S.F.: Detecting Image Near-Duplicate by Stochastic Attributed Relational Graph Matching with Learning. In: ACM MM 2004, pp. 877–884 (2004)
12. Zhao, W., Jiang, Y., Ngo, C.: Keyframe Retrieval by Keypoints: Can Point-to-point Matching Help? In: CIVR 2006, pp. 72–81 (2006)
13. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Kadir, F.S.T., Gool, L.V.: A Comparison of Affine Region Detectors. International Journal of Computer Vision 65(1/2), 43–72 (2005)
14. Mikolajczyk, K., Schmid, C.: A Performance Evaluation of Local Descriptors. IEEE Trans. on Pattern Analysis and Machine Intelligence 27(10), 1615–1630 (2005)
15. Fleck, D., Duric, Z.: Affine Invariant-Based Classification of Inliers and Outliers for Image Matching. In: Kamel, M.S., Campilho, A.C. (eds.) ICIAR 2005. LNCS, vol. 3656, pp. 407–414. Springer, Heidelberg (2005)
16. TREC Video Retrieval Evaluation (TRECVID) (2010),
<http://trecvid.nist.gov/>
17. Ojala, T., Pietikainen, M.: Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. IEEE Trans. on Pattern Analysis and Machine Intelligence 24(7), 791–981 (2002)
18. Fukunaga, K.: Introduction to Statistical Pattern Recognition, 2nd edn. Academic Press, London (1990)
19. <http://www.csse.uwa.edu.au/~pk/Research/MatlabFns/index.html> (2010)

Audio Tag Annotation and Retrieval Using Tag Count Information

Hung-Yi Lo^{1,2}, Shou-De Lin², and Hsin-Min Wang¹

¹ Institute of Information Science, Academia Sinica, Taipei

² Department of Computer Science and Information Engineering,
National Taiwan University, Taipei

Abstract. Audio tags correspond to keywords that people use to describe different aspects of a music clip, such as the genre, mood, and instrumentation. With the explosive growth of digital music available on the Web, automatic audio tagging, which can be used to annotate unknown music or retrieve desirable music, is becoming increasingly important. This can be achieved by training a binary classifier for each tag based on the labeled music data. However, since social tags are usually assigned by people with different levels of musical knowledge, they inevitably contain noisy information. To address the noisy label problem, we propose a novel method that exploits the tag count information. By treating the tag counts as costs, we model the audio tagging problem as a cost-sensitive classification problem. The results of audio tag annotation and retrieval experiments show that the proposed approach outperforms our previous method, which won the MIREX 2009 audio tagging competition.

Keywords: Audio tag annotation, audio tag retrieval, folksonomy, tag count, cost-sensitive learning, cost-sensitive evaluation.

1 Introduction

With the explosive growth of digital music available on the Web, organizing and retrieving desirable music from online music databases is becoming an increasingly important and challenging task. Until recently, most research on music information retrieval (MIR) focused on classifying musical information with respect to the genre, mood, and instrumentation. Social tags, which have played a key role in the development of “Web 2.0” technologies, have become a major source of musical information for music recommendation systems. Music tags are free text labels associated with different aspects of a music clip, like the artist, genre, emotion, mood, and instrumentation [4]. Consequently, music tag classification seems to be a more complete and practical means of categorizing musical information than conventional music classification. Given a music clip, a tagging algorithm can automatically predict tags for the clip based on models trained from music clips with associated tags collected beforehand.

Automatic audio tag annotation has become an increasingly active research topic in recent years [3][5][7][9][10], and it has been one of the evaluation tasks at the Music Information Retrieval Evaluation eXchange (MIREX) since 2008[1]. Our previous method [5], which won the MIREX 2009 audio tag annotation competition, exploited both homogeneous segmentation and classifier ensemble techniques. The runner-up [7] in 2009 viewed the audio tag prediction task as a multi-label classification problem and used a stacked (two-stage) SVM to solve it. Another submission [3] introduced a method called the Codeword Bernoulli Average (CBA) model for tag prediction. It is based on a vector quantized feature representation. In the MIREX 2008 evaluation task, the winning audio tag annotation and retrieval system [10] modeled the feature distribution for each tag with a Gaussian mixture model (GMM). The model's parameters were estimated with the weighted mixture hierarchies expectation maximization algorithm. In this paper, our objective is to improve our 2009 winning method by using tag count information.

The audio tagging task can be evaluated from two perspectives: audio tag annotation and audio tag retrieval, as mentioned in [5]. The audio annotation task is viewed as a binary classification problem of each tag, since a fixed number of tags are given. Each tag classifier determines whether the input audio clip should have a specific tag by outputting a score. The performance can be evaluated in terms of the percentage of tags that are verified correctly, or the area under the receiver operating characteristic curve (AUC) given clip (i.e., the correct tags should receive higher scores). In the audio retrieval task, given a specific tag as a query, the objective is to retrieve the audio clips that correspond to the tag. This can be achieved by using the tag classifier to determine whether, based on the score, each audio clip is relevant to the tag. The clips are then ranked according to the relevance scores, and those with the highest scores are returned to the user. The performance can be evaluated in terms of the tag F-measure or the tag AUC.

Social tagging, also called *folksonomy*, enables users to categorize content collaboratively by using tags. Unlike the classification labels annotated by domain experts, the information provided in social tags may contain *noise* or *errors*. Table I shows some examples of audio clips with associated tags obtained from the MajorMiner [6] website², a web-based game for collecting music tags. Some other details, such as the song's title, album, and artist, are also available. Consider that the tag count indicates the number of users who have annotated the given audio clip with the tag. We believe that tag count information should be considered in automatic audio tagging because the count reflects the confidence degree of the tag. Take the first audio clip from the song *Hi-Fi* as an example. It has been annotated with "drum" nine times, with "electronic" three times and with "beat" twice. Therefore, the tag "drum" is the *major property* of the audio clip. The count also reflects the popularity of the tag, song, artist, and album. To the best of our knowledge, tag count information has not been used in

¹ <http://www.music-ir.org/mirex/2008>

² <http://www.majorminer.org/>

automatic audio tagging previously. In the MIREX audio tagging competition, the tag count is transformed into 1 (with a tag) or 0 (without a tag), by using a threshold. Then, a binary classifier is trained for each tag to make predictions about untagged audio clips. As a result, a tag assigned twice is treated in the same way as a tag assigned hundreds of times. In addition, a tag with a small count may contain noisy information, which would affect the training of the tag classifier. To solve the problem, we propose using the tag count information to train a cost-sensitive classifier that minimizes the training error associated with tag counts.

Another issue is how to evaluate the prediction performance that considers the tag counts. For example, a system that gives a single tag “drum” to the *Hi-Fi* audio clip should be considered better than a system that gives both “electronic” and “beat” to the clip, but misses “drum”. In this paper, we propose two cost-sensitive metrics: a cost-sensitive F-measure and a cost-sensitive AUC. The metrics favor a system that recognizes repeatedly assigned tags.

The contribution of this paper is twofold. First, we employ a cost-sensitive learning approach that treats the tag counts as costs in the audio tag annotation and retrieval task. Second, we propose two cost-sensitive evaluation metrics for the performance evaluation. The results of experiments demonstrate that the proposed cost-sensitive methods outperform their cost-insensitive counterparts in terms of not only the cost-sensitive metrics but also the regular metrics.

The remainder of this paper is organized as follows. In Section 2, we consider some factors that affect the tag counts; and in Section 3, we discuss the proposed cost-sensitive learning methods and cost-sensitive evaluation metrics. In Section 4, we describe the experiments and analyze the results. Section 5 contains some concluding remarks.

Table 1. Some Examples of Audio Clips with Associated Tags Obtained from the MajorMiner Website

Song	Album	Clip Start Time	Artist	Associated Tags (Tag Counts)
Hi-Fi	Head Music	0:00	Suede	drum(9) electronic(3), beat(2)
Universal Traveler	Talkie Walkie	4:00	Air	synth(7), electronic(4), vocal(5) female(4), voice(2), slow(2) ambient(2), soft(3), r&b (3)
Safe	Travis	1:00	The Invisible Band	guitar(5),male(4),pop(4) vocal(3),acoustic(2)
Moritat	Saxophone Colossus	0:50	Sonny Rollins	jazz(9) saxophone(12)
Pacific Heights	Ascension	2:30	Pep Love	male(4), synth(2) hip hop(8), rap(6)
Trouble	The Chillout	3:40	Coldplay	male(6), pop(3), vocal(5) piano(7), voice(3) slow(2), soft(2), r&b(2)

2 Tag Counts

From our study of audio tagging websites, such as Last.fm³ and MajorMiner, we observe that certain factors affect the tag counts:

1. **Consistent Agreement:** Social tags are usually assigned by users (including malicious users) with different levels of musical knowledge and different intentions [4]. Tags may contain a lot of noisy information; however, when a large number of users consider that an audio clip should be associated with a particular tag, i.e., the count of the tag is high, the label information is deemed more reliable. Conversely, if a tag is only assigned to an audio clip once, the annotation is considered untrustworthy. The MajorMiner website does not show such tags because they may contain noise. When training a classifier, using noisy label information can affect the generalization ability of the classifier. Another problem that must be considered is that, sometimes, only a small portion of an audio clip is related to a certain tag. For example, an instrument might only be played in the bridge section of a song. In this case, the count of the tag that corresponds to the instrument will be small.
2. **Tag Bias:** There are several types of audio tags, e.g., genre, instrumentation, mood, locale, and personal usage. Some types (such as genre) are used more often than others (such as personal usage tags like “favorite” and “I own it”); and specific tags (such as “British rock”) are normally used less often than general tags (such as “rock”). In addition, audio tags typically contain many variants [4]. For example, on the Last.fm website “female vocalists” is a common tag, and “female vocals” and “female artists” are variants of it. Figure 1 shows the histogram of the average tag count estimated from MajorMiner data. We observe that the average count of most tags is close to 2.5. The tags with higher average counts are assigned more often than the other tags. The top three most repeatedly assigned tags are “jazz”, “saxophone”, and “rap”; and the tags assigned least repeatedly are “drum machine”, “voice”, and “keyboard.” We believe that repeatedly assigned tags are either more popular or they describe acoustic characteristics that are easier to recognize (e.g., “drum machine” might easily be recognized as “drum”).
3. **Song/Album/Artist Popularity:** Popular songs, albums, and artists usually receive more tags, since people tend to tag music that they like or they are familiar with. However, this is not the case for web-based labeling games because the label flow can be controlled by the game designer. In addition, newly released songs usually receive fewer tags.

Based on the above observations, we formulate the audio tag annotation and retrieval task as a cost-sensitive classification problem. In the next section, we examine the concept of cost-sensitive learning and describe our approach.

³ <http://www.last.fm/>

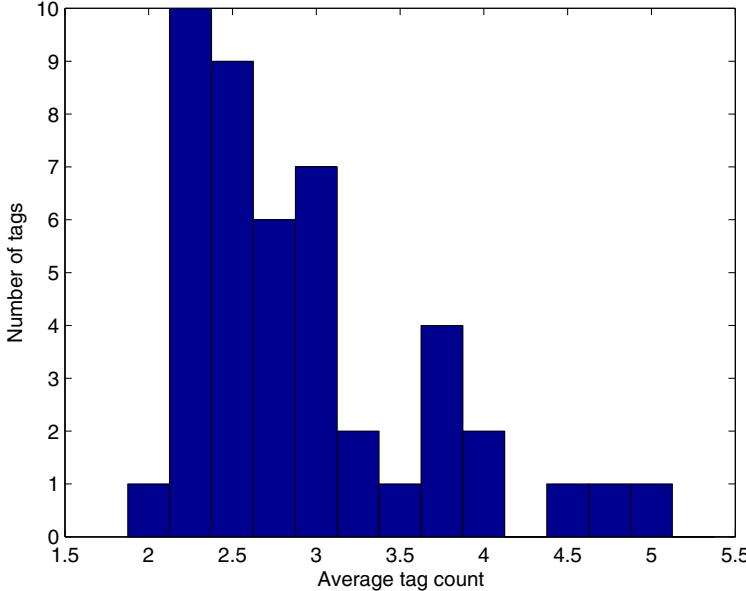


Fig. 1. The histogram of the average tag count

3 Cost-Sensitive Learning

Non-uniform misclassification costs are very common in a variety of pattern recognition applications, such as medical diagnosis, e-mail spam filtering, and business marketing. As a result, several cost-sensitive learning methods have been developed to address the problem, e.g., the modified learning algorithm [8] and the data re-sampling method [11]. Suppose we have a cost-sensitive training set $(\mathbf{x}_i, y_i, c_i)_{i=1}^m$, where $\mathbf{x}_i \in R^n$ is the feature vector of the i -th training sample; $y_i \in \{1, -1\}$ is the class label; and $c_i \subset [0, \infty)$ is the misclassification cost. The goal of cost-sensitive classification is to learn a classifier $f(\mathbf{x})$, that minimizes the expected cost as follows:

$$E[cI(f(\mathbf{x}) \neq y)], \quad (1)$$

where $I(\cdot)$ is an indicator function that yields 1 if its argument is true, and 0 otherwise. The expected cost-insensitive cost is defined as:

$$E[I(f(\mathbf{x}) \neq y)], \quad (2)$$

which is a special case of (1) where all samples have an equal misclassification cost c .

The Paralyzed Veterans of America (PVA) dataset used in the KDD Cup 1998 Competition is a well-known dataset for the cost-sensitive learning problem. It contains information about people who have made at least one prior donation to the PVA, as well as the date and amount of each donation. Participants in the

above competition were asked to train a predictive model that would be used to choose the donors that should be sent a request for a new donation. Since mailing a request to an individual donor involves some cost, the measure of success is the total revenue derived from the mailing campaign. The task is formulated as a cost-sensitive classification problem in [11]. Each instance is associated with a misclassification cost, which is calculated as the donated amount minus the cost of mailing the request for positive instances, and as the cost of mailing the request for negative instances. The predictive model is trained to minimize the total misclassification cost on the training data and is expected to maximize the total revenue on the test data.

Based on the above concept, we formulate the audio tag prediction task as a cost-sensitive classification problem. Specifically, we minimize the total counts of misclassified tags by treating the tag counts as costs. In other words, our goal is to correctly predict the most frequently used tags, such as tags of consistent agreement, popular tags, and the tags for popular songs/albums/artists. For example, consider the first audio clip from the song *Hi-Fi* in Table II and suppose a tag prediction system A only predicts the tag “drum” correctly, while another tag prediction system B predicts two tags “electronic” and “beat” correctly. We consider that system A outperforms system B because the tag “drum” captures the major property of this audio clip.

3.1 Cost-Sensitive Evaluation Metrics

The evaluation metrics used in MIREX 2009, namely, the accuracy, F-measure, tag AUC, and clip AUC, did not consider the costs (i.e., tag counts). Moreover, the class distribution of each binary tag classification problem was imbalanced. For example, in the MajorMiner dataset used at MIREX 2009, out of the forty-five tags, only twelve had more than 10% positive instances. Using accuracy as the evaluation metric biases the system towards the negative class. Since these metrics do not take the costs into account, we propose three cost-sensitive metrics. First, we define the cost-sensitive precision (CSP) and the cost-sensitive recall (CSR) as follows:

$$CSP = \frac{\text{Weighted Sum of TP}}{\text{Weighted Sum of TP} + \text{Weighted Sum of FP}}, \quad (3)$$

$$CSR = \frac{\text{Weighted Sum of TP}}{\text{Weighted Sum of TP} + \text{Weighted Sum of FN}}, \quad (4)$$

where TP, FP, and FN denote the true positive, the false positive, and the false negative, respectively. The weight of each positive instance is assigned as the count of the associated tag. However, assigning a weight to each negative instance is not as straightforward because people do not use negative tags like “non-rock” and “no drum.” Therefore, we assign a uniform cost to negative instances and balance the cost between positive and negative classes, i.e., the total cost of the positive instances is the same as that of the negative instances.

As a result, the expected CSP of a random guess baseline method will be 0.5. Then, we can define cost-sensitive metrics based on CSP and CSR.

The *cost-sensitive F-measure* can be calculated as follows:

$$2 \times \frac{\text{CSP} \times \text{CSR}}{\text{CSP} + \text{CSR}}. \quad (5)$$

The receiver operating characteristic curve (ROC) is a graphical plot of the true positive rate (recall) versus the false positive rate as the decision threshold varies. The area under the ROC curve (AUC) is often used to evaluate a binary classifier's performance on a class-imbalanced dataset. We can modify the AUC to obtain a *cost-sensitive AUC* by replacing the recall metric with CSP. Then, we use the cost-sensitive tag AUC and the cost-sensitive clip AUC to evaluate the audio annotation task and the audio retrieval task, respectively.

3.2 Cost-Sensitive Classification Methods

Support vector machine (SVM) and AdaBoost are two very effective learning algorithms for classification problems. In this subsection, we describe their cost-sensitive versions.

The training process of SVM attempts to maximize the margin and minimize the training error at the same time. The objective function for cost-sensitive SVM is formulated as follows:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^m c_i \xi_i, \\ \text{s.t. } & y_i (\mathbf{w}^T \phi(\mathbf{x}_i) + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, \dots, m, \end{aligned} \quad (6)$$

where ϕ is a function that maps the input data to a higher dimensional space and C is a tuning parameter that exists in the general SVM form. Note that each cost c_i is associated with a corresponding training error term ξ_i .

AdaBoost finds a highly accurate classifier by combining several base classifiers, even though each of them is only moderately accurate. Cost-sensitive AdaBoost [8] maintains a weight vector D_t for the training instances in each iteration and uses a base learner to find a base classifier to minimize the weighted error according to D_t . In each iteration, the weight vector D_t is updated by

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t c_i y_i h_t(\mathbf{x}_i))}{Z_t}, \quad (7)$$

where $h_t(\mathbf{x}_i)$ is the prediction score of the base classifier h_t for instance \mathbf{x}_i ; c_i is the cost of each instance; Z_t is a normalization factor that makes D_{t+1} a distribution; and α_t can be calculated based on different versions of AdaBoost. We use a decision stump as the base learner in this study. The other implementation details are the same as the standard AdaBoost [2].

4 Experiments

The experiments follow our previous setup of the MIREX 2009 extended experiments reported in [5]. We consider forty-five tags, which are associated with

2,472 audio clips downloaded from the website of the MajorMiner game. The duration of each clip is 10 seconds or less.

Given an audio clip, we divide it into several homogeneous segments by using an audio novelty curve [1]. Then, using MIRToolbox 1.1⁴, we extract a 174-dimensional audio feature vector from each segment to reflect various types of musical information, such as the segment’s dynamics, rhythm, timbre, pitch, and tonality. For an audio clip, the prediction score given by a classifier is the average of its constituent segments. The SVM and AdaBoost scores are merged by using either a probability ensemble to annotate an audio clip, or a ranking ensemble to rank all the audio clips according to a tag. For the ranking ensemble, we first rank the prediction scores of individual classifiers independently. Then, a clip’s final score is the average of the rankings derived by the two classifiers. For the probability ensemble, we transform the output score of each component classifier into a probability score with a sigmoid function, and then compute the average of the two probability scores.

4.1 Model Selection and Evaluation

We adopt three-fold cross-validation in the experiments. The audio clips are randomly split into three subsets. In each fold, one subset is selected as the test set and the remaining two subsets serve as the training set. The test set for (outer) cross-validation is not used to determine the classifier’s settings. Instead, we perform inner cross-validation on the held out data from the training set to determine the cost parameter C in SVM and the number of base learners in AdaBoost. Then, we retrain the classifiers with the complete training set and the selected parameters, and perform outer cross-validation on the test set. We use the AUC as the model selection criterion.

To calculate the tag F-measure, we need a threshold to binarize the output score. In the audio retrieval task, we want to retrieve audio clips from an audio database. We assume that each tag’s class has similar probability distributions in the training and testing audio databases; therefore, we set the threshold with the class’s distribution obtained from the training data. In the audio annotation task, we annotate the test audio clips one by one. We set the threshold to 0.5 because the calibrated probability score ranges from 0 to 1.

4.2 Experiment Results

We compare the proposed method, which uses the tag count information, with our MIREX 2009 winning method, which did not use such information. The experiment results in terms of the cost-sensitive metrics and regular metrics are summarized in Tables 2 and 3, respectively. The AdaBoost, SVM, and classifier ensemble are the same as those used in the above winning method. We perform three-fold cross-validation twenty times and calculate the mean and standard deviation of the results to reduce the variance caused by different cross-validation splits. Note

⁴ <http://users.jyu.fi/lartillo/mirtoolbox/>

Table 2. Evaluation results of different classifiers and ensemble methods in terms of cost-sensitive metrics

	Mean ±St.d.	CS Tag AUC	CS Clip AUC	CS F-measure
AdaBoost	0.8055	0.8892	0.4099	
	±0.0027	±0.0011	±0.0052	
	0.8169	0.8967	0.4469	
CS AdaBoost	±0.0023	±0.0005	±0.0081	
SVM	0.8112	0.8957	0.4354	
	±0.0022	±0.0007	±0.0077	
	0.8215	0.9005	0.4593	
CS SVM	±0.0023	±0.0004	±0.0056	
Ensemble	0.8334	0.8979	0.4606	
	±0.0019	±0.0007	±0.0067	
	0.8356	0.9032	0.4808	
CS Ensemble	±0.0018	±0.0006	±0.0072	

Table 3. Evaluation results of different classifiers and ensemble methods in terms regular (cost-insensitive) metrics

	Mean ±St.d.	Tag AUC	Clip AUC	F-measure
AdaBoost	0.7941	0.8773	0.3018	
	±0.0027	±0.0011	±0.0035	
	0.8050	0.8854	0.3216	
CS AdaBoost	±0.0023	±0.0005	±0.0049	
SVM	0.7992	0.8837	0.3226	
	±0.0021	±0.0007	±0.0053	
	0.8106	0.8894	0.3299	
CS SVM	±0.0021	±0.0004	±0.0037	
Ensemble	0.8221	0.8859	0.3386	
	±0.0018	±0.0007	±0.0046	
	0.8247	0.8921	0.3442	
CS Ensemble	±0.0017	±0.0005	±0.0046	

that the tag AUC and F-measure are more suitable for the audio retrieval task, while the clip AUC is more suitable for the audio annotation task.

The results in Table 2 demonstrate the effectiveness of using the tag counts to train a cost-sensitive classifier. The cost-sensitive methods outperform their cost-insensitive counterparts. The improvement in the cost-sensitive F-measure is the most significant: 3.7% for CS AdaBoost versus AdaBoost and 2.4% for CS SVM versus SVM. In addition, SVM slightly outperforms AdaBoost, and the ensemble methods outperform the classifiers using SVM or AdaBoost alone.

Table 3 compares the results of different methods in terms of the regular (cost-insensitive) evaluation metrics. Interestingly, the cost-sensitive methods outperform their cost-insensitive counterparts in terms of the regular metrics. Recall that tags with smaller counts may contain *noisy labeling information*. By viewing the tag counts as costs, the cost-sensitive learning method can ignore the noisy information by giving a smaller penalty (cost), and thereby train a more accurate classifier.

5 Conclusion

We have proposed a novel method for exploiting tag count information in audio tagging tasks, and discussed several factors that affect the counts of tags assigned to an audio clip. Among the factors, we believe that consistent agreement is the most important issue, since noisy labeling information in tags assigned less frequently impacts the training of a tag classifier. We formulate the audio tag prediction task as a cost-sensitive classification problem in order to minimize the misclassified tag counts. In addition, we present cost-sensitive versions of several regular evaluation metrics. The proposed cost-sensitive methods outperform their cost-insensitive counterparts in terms of not only the cost-sensitive evaluation metrics but also the regular evaluation metrics.

We have realized that the audio tagging task can also be formulated as a multi-label classification problem. In our future work, we will develop a cost-sensitive multi-label learning algorithm for the task. We will also evaluate our methods on more multimedia tagging tasks.

References

1. Foote, J., Cooper, M.: Media segmentation using self-similarity decomposition. In: SPIE (2003)
2. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1) (1997)
3. Hoffman, M., Blei, D., Cook, P.: Easy as cba: A simple probabilistic model for tagging music. In: ISMIR (2009)
4. Lamere, P.: Social tagging and music information retrieval. Journal of New Music Research 37(2), 101–114 (2008)
5. Lo, H.Y., Wang, J.C., Wang, H.M.: Homogeneous segmentation and classifier ensemble for audio tag annotation and retrieval. In: ICME (2010)
6. Mandel, M.I., Ellis, D.P.W.: A web-based game for collecting music metadata. In: ISMIR (2007)
7. Ness, S., Theocharis, A., Tzanetakis, L.G.M., Improving, G.: automatic msic tag annotation using stacked generalization of probabilistic svm outputs. In: ACM MM (2009)
8. Sun, Y., Kamel, M.S., Wong, A.K.C., Wang, Y.: Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition 40(12), 3358–3378 (2007)

9. Tingle, D., Kim, Y., Turnbull, D.: Exploring automatic music annotation with “acoustically-objective” tags. In: MIR (2010)
10. Turnbull, D., Barrington, L., Torres, D., Lanckriet, G.: Semantic annotation and retrieval of music and sound effects. IEEE Trans. on Audio, Speech, and Language Processing 16, 467–476 (2008)
11. Zadrozny, B., Langford, J., Abe, N.: Cost-sensitive learning by cost-proportionate example weighting. In: ICDM (2003)

Similarity Measurement for Animation Movies

Alexandre Benoit¹, Madalina Ciobotaru¹, Patrick Lambert¹,
and Bogdan Ionescu²

¹ LISTIC - Universite de Savoie 74940, Annecy le Vieux, France

² LAPI, University "Politehnica" Bucharest, 061071, Romania

`alexandre.benoit@univ-savoie.fr, patrick.lambert@univ-savoie.fr,
BIonescu@alpha.imag.pub.ro`

Abstract. When considering the quantity of multimedia content that people and professionals accumulate day by day on their storage devices, the necessity of appropriate intelligent tools for searching or navigating, becomes an issue. Nevertheless, the richness of such media is difficult to handle with today's video analysis algorithm. In this context, we propose a similarity measure dedicated to animation movies. This measure is based on the fuzzy fusion of low level descriptors. We focus on the use of a Choquet Integral based fuzzy approach which is proved to be a good solution to take into account complementarity or conflict between fused data and so to model a human like similarity measure. Subjective tests with human observers have been carried out to validate the model.

Keywords: Movie similarity measure, fuzzy fusion, animation movies, video sequences.

1 Introduction

In many domains, there is a rapidly growing amount of image and video information which are published and stored. This phenomenon creates the need for an efficient way of searching and navigating through the content. In the literature, there is a lot of works on image, video, or more generally multimedia indexation and retrieval [1] [2]. Among these works, a specific attention is dedicated to content-based retrieval using generally the traditional "query by example". In such an approach, the system looks for a response which is similar to the query, which supposes an automatic similarity measure. In the case of still images, despite the subjective aspect of this measure, there is a very large variety of methods able to measure the similarity between two images [7]. For this purpose, commonly used features are colors, textures, shapes, and more generally "bag of features". In the case of videos, the problem is much more complex. This is mainly due to the much larger amount of data and to the even more subjective definition of similarity between two videos. A first solution consists in using image similarity techniques applied on key images representing the videos. The drawback is that it does not take into account the motion or temporal information included in the videos. An alternative solution is proposed in [8] where authors compute the average distance of all the corresponding

frames, which may be time-consuming. On the other hand, many works have been done in the specific case of “copy detection”, proposed in the TrecVid challenge (<http://www-nlpir.nist.gov/projects/tv2008/>), for searching a specific sub-shot (for instance looking for services in a tennis match) [9] or for automatic genre classification [3], [4], [5], [6].

Only a very few works attempt to measure a video similarity from a global point of view. In [10], Cheung proposes to compare two videos by measuring the similarity between their compact color signatures. Peng [11] uses a hierarchical approach based on graph matching. These two approaches are used on web video sequences. In [12], Lienhart asks three basic questions: at what temporal duration can video sequences be compared? Given some images or video features, what are the requirements on its distance measure and how can it be easily transformed into the visual similarity desired by the inquirer? How can video sequences be compared at different levels?

In this paper, we propose a solution to compare short animation movies from a global point of view. The aim is to provide intuitive and human like animation movie comparison which could help database navigation. The proposed approach is based on a “ground truth” provided by a human evaluation campaign requiring movie pair comparisons. As the time necessary for this evaluation is exponentially depending on the number of involved movies, in a first approach, the database size is limited to 51 movies. This first attempt will also allow context dedicated test protocols setup to be enhanced.

The paper is organized as follow: section 2 presents the specific context of animation movies. Section 3 gives a detailed presentation of the proposed approach. Experimental results are discussed in Section 4. Finally, section 5 proposes some conclusions and perspectives.

2 Animation Film Context

The “International Animated Film Festival”, managed by CITIA, is an yearly festival (<http://www.citia.info>), which takes place in Annecy (France), since 1960. It the most important event in the worldwide animated movie entertainment community. In 2010 festival, 70 countries were present in the program festival. A few years ago, CITIA had the initiative of digitizing most of the movies, to compose one of the first digital animated movie database. Today, this database covers more than 31.000 movie titles. At the moment, the existing indexing tools only use the textual information provided mainly by movie authors, which in many cases does not totally apply to the rich artistic content of the animated movies. So there is a real need for complementary tools which would involve the visual information to improve research or navigation.

Animation movies from CITIA are different from classical animation movies (i.e. cartoons) or from natural movies, in many aspects : their duration is generally small (typically 10 mn); they are mainly fiction or abstract movies, therefore everything is possible; characters, if any, can take any shape or color, thus face recognition and skin detection techniques are useless; there is a very large variety



Fig. 1. Various animation techniques (from left to right and top to bottom): paper drawing, object animation, 3D synthesis, glass painting, plasticine modeling and color salts

of animation techniques; colors are selected and mixed by the artists to express particular feelings, therefore each movie has a specific color signature. Finally, content is very varied, and it is impossible to use a priori information as it is classically done when analyzing informative videos or sport videos. Figure 1 shows examples of animation movies using different techniques.

3 The Proposed Approach

3.1 General Description

The proposed approach is based on two main steps as shown in figure 2. The first one is an evaluation of the movie similarities provided by human observers. This step will be detailed in the following sub-section. Its aim is to get a human similarity measure for each pair of movies within a set of 51 movies. The second step is an automatic evaluation tool giving similarity measures which are designed to approach human evaluations. This step is composed of two main parts. The first one is the extraction of low level features from image sequences. The second one is a fusion process aggregating, for two movies, the differences between corresponding features. The fusion is tuned according to each human evaluation and is designed to fit these human measures.

3.2 Evaluation by Human Observers

51 sequences coming from CITIA have been chosen. This dataset allows 1275 movie comparisons. Subjective tests with human observers are generally carried out following strictly controlled protocols such as ITU-BT500.11 recommendations [13]. It includes restrictions such as limited duration experiments length (typically 30min), controlled environment, etc. However, when considering involved films length and the exploration study context, standard recommendations cannot be entirely applied. Furthermore, as these studies are explorative,

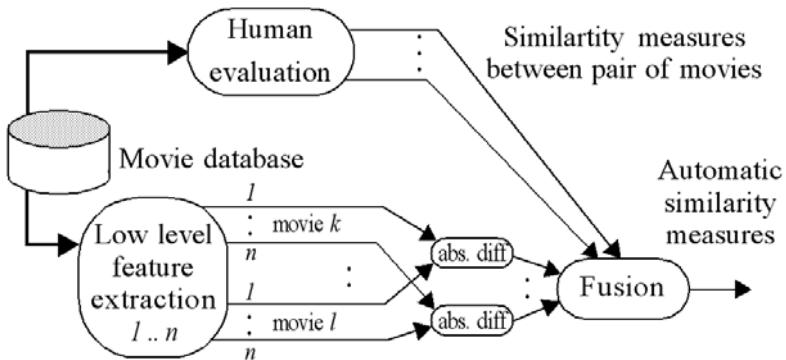


Fig. 2. General overview of the proposed approach

Observer A opinions Observer B opinions Observer C opinions

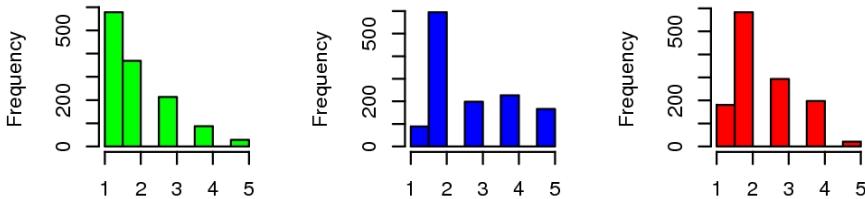


Fig. 3. Observer opinion histograms

observer behaviors are unknown and constrained us to analyse each observer independently. Then, observers took part to 12 hours experiments divided in 40 min sessions spreaded on 18 days, in one month. All sessions took place in the same place and similar conditions (light, time, hardware) and sequences pairs where seen in random order. Considering the huge experiment length, only three persons took part in subjective tests. All were non experts in animation movies techniques and culture. Ages vary within range 22 and 32 years old. First results are intended to benefit to further studies with refined test protocols, a larger test set and more observers. The small amount of persons involved in these tests is a real problem. However this first attempt will be used to design larger experiments with lighter human observer procedures.

We conducted tests following Absolute Category Rating (ACR) methodology also called Single Stimulus Method. The evaluation step consisted in evaluating the distance between the considered two sequences on a 5 category scale: [1: very different, 2: different, 3: median, 4: similar and 5: very similar]. Obtained opinions histograms are shown on figure 3.

This preliminary study involves observers with different opinions statistics. Observer A chose mostly the 'Very different' (45%) category while other categories weight decrease gradually when similarity grows. Observer B and C mostly respond 'Different' (47%) but other categories are more homogeneous (around 13%).

Table 1. Observers opinions comparison

Compared observers	CC	RMSE
A v.s. B	0.48	13%
A v.s. C	0.50	8%
B v.s. C	0.73	5%

However, observer A and C gave few 'Very Similar' category (less than 5%). Table 1 allows inter-observer opinion analysis. Root Mean Square Error (RMSE) is used to measure the average similarity between observers. Linear correlation coefficient (CC) reports the monotony in the relation between observers evaluations [16]. These two measures are generally used for subjective evaluation analysis [17]. RMSE remains lower than 15% which expect global scores to be close. However CC is medium which means that observer measure models are slightly different and probably correlated in a non-linear way. This fact encourages us to investigate single person similarity measurement modelling. In a further study, a larger scale experiment will be necessary to identify mean opinion rules.

3.3 Feature Extraction

Different low-level features are extracted from the animation movies (previous works [14]). The specificity of these features is that they give a global characterization of each movie. This type of description is relevant for animation movies which duration is small (typically 10 minutes) and would probably fail in the case of long duration movies or videos. Used features are mainly related to the global color distribution and to the movie structure in terms of video transition distribution : the movie color distribution provides us with detailed information regarding the movies artistry content while the movie structure, thus the shot distribution, provides us with information on the movie action content (temporal segmentation is detailed in [14]).

In order to shorten this part, we only present the five most efficient features that we have selected in the following: four characterizing the color distribution and one related to the transition distribution : the "dark color ratio", the "weak color ratio", the "warm color ratio", the "color variation ratio" and the "mean cut speed".

To get the color features, a global normalized color histogram h_{seq} is computed. This histogram is a weighted average of the color histograms obtained for each shot, the weighting coefficients being related to the shot duration. Color histogram of each shot is computed using the webmaster palette, a predefined color palette of 216 colors (<http://www.visibone.com/colorlab>). This palette is interesting for two reasons. First, it is a good compromise between total number of colors and color wealth. Second, it is associated to a color naming system which is essential to define relevant color features in the following way:

The “dark color ratio” f_{dark} is the proportion of colors which use “dark”, “obscure” or “black” in their name:

$$f_{dark} = \sum_{c=1}^{216} h_{seq}(c) | \text{Name}(c) \text{ includes “dark”, “obscure” or “black”}$$

where h_{seq} denotes the movie histogram and c denotes the webmaster palette colors.

In the same way, the “weak color ratio” (resp. the “warm color ratio”) is defined by the proportion of colors which use “dull”, or “weak” (resp. “yellow”, “orange”, “red”, “violet”, “magenta”, “pink”, or “spring”) in their name.

The “color variation ratio” f_{dark} is the proportion of colors c , among the 216 colors, for which $h_{seq}(c) > 1\%$ (empirically chosen threshold). This characteristic is related to the color wealth used in the movie.

The last feature, the “mean cut speed”, is an average of the number of transitions between shots calculated within a sliding window of 5s. This last feature is normalized (using minimum and maximum values over a wide set of movies) to get a value in the $[0,1]$ interval as all the color features.

So, all the used features are within the range $[0,1]$ which ensures the commensurability property necessary for the next step fusion.

3.4 Fusion of Feature Differences

Let us denote f_i^k the value of feature i ($i = 1..n$) for movie k . To compare two movies k and l , we have a set of n differences $\delta_i^{k,l} = |f_i^k - f_i^l|$ which have to be fused to get one similarity measure between the two movies. As each feature value is within $[0,1]$, it can be noted that the differences will also be within $[0,1]$.

A classical way consists in using a weighted average where the weights are adjusted thanks to a learning step using the human evaluation. However, this solution doesn't take into account the eventual complementarities or conflicts between the differences $\delta_i^{k,l}$. So we propose to use another approach based on the Choquet integral which belongs to the fuzzy integral family [15]. This approach is able to consider the interactions between differences in addition to the classical weighted average. In this paper we use the 2-additive Choquet integral given by:

$$\text{sim}_{k,l} = \sum_{i=1}^n \alpha_i \cdot \delta_i^{k,l} - \frac{1}{2} \cdot \sum_{i=1}^n \sum_{j=i+1}^n \beta_{i,j} \cdot |\delta_i^{k,l} - \delta_j^{k,l}|$$

where $\beta_{i,j}$ denotes the weight of the interaction between differences $\delta_i^{k,l}$ and $\delta_j^{k,l}$.

All the weights α_i and $\beta_{i,j}$ are tuned in a learning step. This learning step uses the human evaluation and adjusts all the weights to get an automatic measure as close as possible to the human evaluation. As the original human evaluations remain within scale 1-5 (see section B.2), they are re-scaled within range $[0,1]$ (0 meaning very different while 1 expresses high similarity). This learning step is achieved using least square identification with the help of Kappalab [15].

So, $\text{sim}_{k,l}$ is the final automatic similarity measure. This measure is within the range $[0,1]$, the movies k and l being as similar as the measure is close to 1.

Table 2. Distance measure model performances

	RMSE/Observers	A	B	C
color features	Weighted sum	33.0%	18.0%	21.2%
	Choquet	21.9%	12.9%	14.6%
color + action	Weighted sum	33.0%	17.7%	21.1%
	Choquet	22.7%	12.8%	14.3%

4 Experimental Results

Performance tests have been carried out using cross validation methodology. Human observations and the related objective measures are separated into a training set on which Choquet integral coefficients are learned and a test set on which performances are analyzed (training set = 2/3 test set = 1/3). Such experiments are performed 3 times in order to involve all the dataset into learning and test steps and we present the mean results. It is important to note that in these first experiments, Choquet coefficients are learned independently for each human observer.

First, performances are measured using only the 4 colors features. In a second analysis the action feature is added. This last experiment is intended to study the impact of animation movie action in the similarity measure. In both cases, we compare results obtained with the classical weighted average to those obtained with Choquet integral. Table 2 presents relative RMSE between the automatic similarity measure $sim_{k,l}$ and the re-scaled ground truth for each observer. It can be noted that, as Choquet integral is non linear, RMSE is a more relevant indicator than the linear correlation coefficient. From a general point of view, model identification is significantly improved when Choquet integral is used instead of the classical weighted average. Choquet allows from 5% to 11% performance improvement respectively for observer B and A. Indeed, interactions between the considered criteria exist and refine the similarity metric. Considering only Choquet based, observer A is the most difficult to model, RMSE being close to 22%, while observers B and C opinions are modelled with less than 13% and 15% error. This shows that using the same limited number of low level objective measures, it is possible to correctly model non experts observers opinions.

When the action criteria is considered, performances are not significantly enhanced (degraded by 0.8% for observer A (RMSE being 22.7%)) while being improved by less than 0.3% for B and C). So, one can say that action information does not impact significantly on the distance measure. Furthermore, when considering the Choquet coefficients, it becomes possible to identify weights and relation strength between the considered criteria. Coefficient analysis validates the preliminary conclusions: observer A is not influenced by action (action weight is null), while a slight impact is noticed for observers B and C (action weight close to 0.1) which explains the small performance improvement. It is also interesting to see that some interactions remain significant for all observers. Particularly, there is a negative interaction weight (-0.3) between (DarkColorRatio) and (WarmColorRatio) which indicates redundancy between these two characteristics.

5 Conclusion

We proposed an attempt to measure similarity between animation video sequences. A metric based on low level movie characteristics have been proposed. The semantic gap between high level human evaluation and the considered low level measures is considered by a fusion process using the Choquet integral. It has been shown that Choquet integral significantly improves results compared to the classical weighted average. A validation experiment has been conducted using human observers and shows the feasibility of the approach. Nevertheless, an issue regarding subjective test protocols has been pointed out due to test durations.

This preliminary study showed that, for each person, it is possible to identify a fusion algorithm which would model his similarity measure with correct performances. It seems that it is possible to use the same low level criteria to model different persons opinions. However this work must be extended to investigate this approach on a larger audience.

Acknowledgements

This work was partially supported under Rhône-Alpes region, Research Cluster ISLE - LIMA project and under FSE - European Structural Funds project EXCEL POSDRU/89/1.5/S/62557 (2010-2013).

References

1. Worring, M., Schreiber, G.: Semantic image and video indexing in broad domains. *IEEE Trans. on Multimedia* 9, 909–911 (2007)
2. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Comput. Commun. Appl.* 2(1), 1–19 (2006)
3. Brezeale, D., Cook, D.J.: Automatic Video Classification: A Survey of the Literature. *IEEE Trans. on Systems, Man and Cybernetics, Part C: Applications and Reviews* 38(3), 416–430 (2008)
4. Roach, M.J., Mason, J.S.D.: Video Genre Classification using Dynamics. In: *IEEE Inter. Conf. on Acoustics, Speech and Signal Processing*, Utah, USA, pp. 1557–1560 (2001)
5. Yuan, X., Lai, W., Mei, T., Hua, X.S., Wu, X.Q., Li, S.: Automatic video genre categorization using hierarchical SVM. In: *IEEE Inter. Conf. on Image Processing*, Atlanta, USA, pp. 2905–2908 (2006)
6. Montagnuolo, M., Messina, A.: Parallel Neural Networks for Multimodal Video Genre Classification. *Multimedia Tools and Applications* 41(1), 125–159 (2009)
7. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys* 40(1) (2008)
8. Wu, Y., Zhuang, Y., Pan, Y.: Contentbased video similarity model. In: *Proceedings of the Eighth ACM International Conference on Multimedia*, MULTIMEDIA 2000, New York, NY, USA, pp. 465–467 (2000)

9. Zhao, Y., Zhou, X.-z., Tang, G.: Audiovisual integration for racquet sports video retrieval. In: Li, X., Zaïane, O.R., Li, Z.-h. (eds.) ADMA 2006. LNCS (LNAI), vol. 4093, pp. 673–680. Springer, Heidelberg (2006)
10. Cheung, S.S., Zakhori, A.: Efficient video similarity measurement with video signature. In: ICIP, pp. 621–624 (2002)
11. Peng, Y., Ngo, C.-W.: Clip-based similarity measure for hierarchical video retrieval. In: Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval, MIR 2004, New York, NY, USA, pp. 53–60 (2004)
12. Lienhart, R., Effelsberg, W., Jain, R.: Visualgrep: A systematic method to compare and retrieve video sequences. *Multimedia Tools Appl.* 10(1), 47–72 (1999)
13. ITU, Methodology for the subjective assessment of the quality of television pictures, Recommendation itur bt.500-11, International Telecommunication Union - Radiocommunication Sector (2002)
14. Ionescu, B., Coquin, D., Lambert, P., Buzuloiu, V.: A fuzzy color-based approach for understanding animated movies content in the indexing task. *EURASIP Journal on Image and Video Processing* (2008)
15. Grabisch, M., Kojadinovic, I., Meyer, P.: A review of methods for capacity identification in choquet integral based multi-attribute utility theory: Applications of the kappalab r package. *European Journal of Operational Research* 186(2), 766–785 (2008)
16. Pearson, K.: Mathematical contributions to the theory of evolution. III. regression, heredity, and panmixia. *Royal Society of London Philosophical Transactions Series A* 187, 253–318 (1896)
17. VQEG, Final report from the video quality experts group on the validation of objective models of video quality assessment, Technical Report, Video Quality Experts Group (2000)

A Feature Sequence Kernel for Video Concept Classification

Werner Bailer

JOANNEUM RESEARCH Forschungsgesellschaft mbH
DIGITAL – Institute of Information and Communication Technologies
Steyrergasse 17, 8010 Graz, Austria
werner.bailer@joanneum.at

Abstract. Kernel methods such as Support Vector Machines are widely applied to classification problems, including concept detection in video. Nonetheless issues like modeling specific distance functions of feature descriptors or the temporal sequence of features in the kernel have received comparatively little attention in multimedia research. We review work on kernels for commonly used MPEG-7 visual features and propose a kernel for matching temporal sequences of these features. The sequence kernel is based on ideas from string matching, but does not require discretization of the input feature vectors and deals with partial matches and gaps. Evaluation on the TRECVID 2007 high-level feature extraction data set shows that the sequence kernel clearly outperforms the radial basis function (RBF) kernel and the MPEG-7 visual feature kernels using only single key frames.

1 Introduction

Kernel methods, most notably Support Vector Machines (SVMs), have been widely applied to classification problems, also due to the availability of toolkits such as LibSVM [4]. SVM based classifiers are also commonly used for concept classification based on visual features. If we look at the TRECVID [16] 2009 High-Level Feature Extraction (HLFE) Task, which is one of the most important benchmarks for concept classification in video data, all but 3 of the 42 submitters report the use of an SVM variant in some part of their approach (e.g. for classification based on low-level features or for fusion) [1]. Most of the groups use some low-level features which require other distances than the Euclidean distance between feature vectors, e.g. some of the MPEG-7 visual descriptors [12] or variants of histograms. But only about half of these groups mention the use of specific kernels for these features, while most seem to use the commonly applied radial basis function (RBF) kernel.

Despite the wide use of MPEG-7 visual features in the research community there is remarkably little work on defining kernels that appropriately model the proposed distance functions. A kernel combining different MPEG-7 features and considering the appropriate distance functions has been proposed in [7,6] and

has shown to perform better on a small still image data set. We use this work and apply it to concept classification in video data on a larger data set.

The second aspect that is often neglected is the temporal dimension of concepts. For example, some of the concepts in the TRECVIDE HLFE task, such as “people marching” imply a temporal dimension. Nonetheless concept classification still often relies on visual features from key frames that fail to capture this dimension. One reason for this might be the problem of plugging temporal features appropriately into a classification framework, given the fact that the segments to be matched differ in length and incomplete matches need to be supported.

In [18] a sequence alignment kernel for matching trajectories is proposed. Segments of trajectories are labeled to obtain both a symbol and a feature sequence. A pair-Hidden Markov Model (pair-HMM) is used to align the symbol sequences and to determine the joint probability distribution of the sequences. Gaps in the alignment are padded with zeros in the feature sequence and a RBF kernel is applied to the feature sequence. The sequence alignment kernel is defined as the tensor product of the kernels evaluated on the symbol and feature sequences. The authors of [14] propose a temporal kernel for concept classification. They train a HMM for each concept and define a distance between HMMs based on the Kullback-Leibler divergence. This distance is plugged into a Gaussian kernel. The approach outperforms classifiers using the same features without temporal information on 30 out of 39 of the LSCOM-lite [13] concepts.

The pyramid match kernel [8] has been proposed for matching sets of features of possibly different cardinality by determining their largest intersection. The approach has been extended to spatiotemporal matching in [5]. It has been applied to discrete sets of clustered SIFT and optical flow features. Another temporal matching method based on the pyramid match kernel is described in [19]. Temporally constrained hierarchical agglomerative clustering is performed to build a structure of temporal segments. The similarity between segments is determined using the Earth Mover’s distance and the pyramid kernel is applied to the similarities on the different hierarchy levels.

The authors of [20] use the Levenshtein distance between sequences of clustered local descriptors for classification of still images. Recently a kernel for event classification based on sequences of histograms of visual words has been proposed [3]. The authors consider different similarity measures between the histograms and use them instead of symbol equality in the Needleman-Wunsch distance. The result of string matching is then plugged into a Gaussian kernel. Relative margin support tensor machines [10] have been recently proposed for spatiotemporal classification problems such as gait and action recognition. They are defined as extensions of SVMs for spatiotemporal data, however, currently they are restricted to data of the same temporal length and do not support flexible alignment.

Some of the proposed approaches require discretizing feature values (e.g. by clustering into a code book) or need a model of the temporal sequence of a class such as a HMM. A drawback of the pyramid matching based approaches is the

fixed spatiotemporal grid that does not allow for temporal offsets or gaps. We propose a method that can use the low-level feature vectors without discretizing, as well as supporting partial matches with gaps. It thus shares some properties with the approach in [3], but it does not only define the kernel value by the longest common subsequence but based on all matching subsequences. In addition it considers the appropriate matching of the widely used MPEG-7 visual features and integrates them into a sequence matching kernel.

The rest of this paper is organized as follows. In Section 2 we review the existing definition of an appropriate kernel for a set of MPEG-7 visual features and use it as a basis for proposing a kernel for a sequence of such feature vectors in Section 3. In Section 4 we report about the evaluation of the MPEG-7 kernel and the sequence kernel on the TRECVID HLFE 2007 data set. Section 5 concludes the paper.

2 A Kernel for MPEG-7 Visual Features

We use four of the MPEG-7 visual features defined in [9] with the distance measures proposed in [12]: Color Layout (CLD), Dominant Color (DCD), Color Structure (CSD) and Edge Histogram (EHD). In this section we review the definition of the kernel proposed in [7]. The joint kernel for our set of features is defined as

$$\kappa_{mpeg7}(x, x') = \prod_{i \in \{cld, dcd, csd, ehd\}} \exp(-\bar{w}_i \kappa_i(x, x')) \quad (1)$$

$\kappa_i(\cdot)$ is a specific kernel for each feature as discussed below. The feature weights w_i are defined as

$$w_i(T) = \frac{\text{var}(\{d_i(x_i^-, y_i^-) | \forall x^-, y^- \in T^-\})}{\text{var}(\{d_i(x_i^+, y_i^+) | \forall x^+, y^+ \in T^+\})}, \quad (2)$$

where x^+ (x^-) denotes a positive (negative) sample in the training set T and $d_i(\cdot)$ is the distance function for feature i . The weights are thus defined as the ratio of the variances of the feature distances among the negative and positive samples. The weights for the individual features are then normalized to obtain $\bar{w}_i = \frac{w_i}{\sum_{j \in \{cld, dcd, csd, ehd\}} w_j}$. In contrast to [7] we calculate the weights in advance and not iteratively during training.

As proposed in [7], the Laplace kernel is used for CSD and EHD. For EHD the appropriate combination of local, global and semiglobal histograms as described in [12] is used. A RBF kernel is used for CLD with appropriate weighting of the distances of the different coefficients.

The DCD descriptor consists of $N \leq 8$ tuples (p_i, \mathbf{c}_i) , where p_i is the percentage of the i^{th} color cluster in the image, and $\mathbf{c}_i = (l_i, u_i, v_i)$ is the center of the color cluster in Luv color space. The distance proposed in [12] is defined as

$$d_{dcd}(x, x') = \sqrt{\sum_{i=1}^N \sum_{j=1}^{N'} (p_i^2 a_{ii} + p_j'^2 a_{jj} - 2 p_i p_j' a_{ij})} \quad (3)$$

where

$$a_{ij} = \begin{cases} 1 - \frac{|\mathbf{c}_i - \mathbf{c}'_j|}{d_{max}} & \text{if } |\mathbf{c}_i - \mathbf{c}'_j| \leq d_{max}, \\ 0, & \text{otherwise,} \end{cases} \quad (4)$$

d_{max} being the maximum distance between clusters. The DCD distance function compares pairwise all combinations of the up to eight colors in the descriptor, thus a kernel performing element by element matching (such as a RBF kernel would do) cannot be applied. In addition, the color distances need to be weighted by the percentages p_i of the colors.

We use the DCD kernel defined in [6]. We define a M -dimensional feature space, where $M = N_{max}^2$ and $N_{max} = \max(N, N')$, i.e. the maximum of the number of colors in both descriptors. Then the percentage values are mapped to vectors

$$\begin{aligned} u &= (\underbrace{p_1, \dots, p_1}_{\times N'}, \dots, \underbrace{p_N, \dots, p_N}_{\times N'}), \\ u' &= (\underbrace{p'_1, \dots, p'_{N'}}_{\times N}). \end{aligned} \quad (5)$$

If $N \neq N'$ the vectors are padded with zeros. We further define a matrix $A = \text{diag}(a_{ij})$. The kernel modeling the distance function is then given as

$$\kappa_{dcg}(x, x') = \exp \left(-\sqrt{(u - u') A (u - u')^T} \right). \quad (6)$$

3 A Kernel for Sequences of Feature Vectors

In order to deal with sequences of feature vectors, we define a kernel based on the distance of subsequences of feature vectors. Subsequences need not be contiguous, but they are ordered. In our work we use the kernel described in Section 2 to match the feature vectors of elements in the two sequences, but the proposed sequence kernel is general enough to plug in any appropriate kernel for the feature vectors of the elements.

The sequence kernel proposed here is based on the idea of the *all-subsequences kernel* [15] for strings. This kernel is defined as

$$\kappa(x, x') = \sum_{s \in \Sigma^*} \phi_s(x) \phi_s(x'), \quad (7)$$

where s denotes a string from the possible set of strings Σ^* , and $\phi_s(x)$ counts the number of times s occurs as a subsequence of x . Clearly, $\phi_s(x) \phi_s(x')$ is only non-zero, if s is a subsequence of both x and x' . We thus only need to consider the set of *common* subsequences $\Sigma_{(x, x')}^*$, and specifically only those that are not a subsequence of another sequence in this set (as they can be counted with the sequence containing them as described below):

$$\Sigma_{(x, x')}^+ = \left\{ s \in \Sigma_{(x, x')}^* \mid \nexists s' \in \Sigma_{(x, x')}^*, s' \neq s, \text{s.t. } s \notin \Sigma_{s'}^* \right\}. \quad (8)$$

In our case, the symbols in x and s are feature vectors of the key frames of the videos to be matched. The feature vectors contain real-valued entries, thus

we cannot define a subsequence by equality of symbols and $\Sigma_{(x,x')}^+$ is not a finite set. The problem of applying string matching methods to real- and vector-valued data has been studied in literature (see e.g. [2] for an overview of relevant work). In the following we use the LCSS distance measure proposed in [2], but in principle any measure for comparing sequences of feature vectors could be used.

This distance measure determines a set of common subsequences of the input video. We do not enforce the minimum length of match and gap constraints proposed in [2], but we use all matching subsequences, i.e. also one frame matches will be considered and gaps may be arbitrarily large. The reason is that we apply the kernel to key frames of shots, and there are in many cases only one or a few key frames per shot.

The LCSS based distance measure defines a subsequence based on the evaluation of the distance between the feature vectors of the elements in the sequence. The distance measure is defined recursively, increasing the match value for each pair of elements with a distance below a threshold θ . Following the idea of a weighted combination of kernels for different lengths of matching subsequences (cf. [15]), we can define our kernel as

$$\kappa_{seq}(x, x') = \frac{1}{\min(n_x, n_{x'})} \sum_{s \in \Sigma_{(x,x')}^+} \text{LCSS}_s(x, x'), \quad (9)$$

where n_x denotes the length of the input sequence x . We plug a kernel $\kappa_f(\cdot)$ for matching the feature vectors of two elements into the recursive distance measure definition from [2] to obtain

$$\begin{aligned} \text{LCSS}(x, x') = \\ \begin{cases} 0, & \text{if } n_x = 0 \vee n_{x'} = 0, \\ \kappa_f(x_{n_x}, x'_{n_{x'}}) + \text{LCSS}(\text{Head}(x), \text{Head}(x')), & \text{if } \kappa_f(x_{n_x}, x'_{n_{x'}}) \geq \theta, \\ \max(\text{LCSS}(\text{Head}(x), x'), \text{LCSS}(x, \text{Head}(x'))) & \text{otherwise,} \end{cases} \end{aligned} \quad (10)$$

where $\text{Head}(x) = (x_1, \dots, x_{n_x-1})$. Note that due to using the kernel function instead of a distance, the constraint has become $\geq \theta$ instead of $< \theta$. We use $\kappa_f(\cdot) = \kappa_{mpeg7}(\cdot)$ as defined above, but any kernel appropriate for the feature vectors can be used.

We still need to define $\text{LCSS}_s(x, x')$, which yields the length of common sequences $s \in \Sigma_{(x,x')}^+$ instead of the length of only the *longest* common sequence $\text{LCSS}(x, x')$. The usual dynamic programming implementation of string matching builds a $n_x \times n_{x'}$ matrix containing the length of the longest match up to this point in each cell. Starting from the right column and bottom line of this matrix and performing backtracking allows recovering the common matching subsequences. We stop backtracking once we encounter an element already visited during previous backtracking steps. Thus,

$$\begin{aligned} \sum_{s \in \Sigma_{(x,x')}^+} \text{LCSS}_s(x, x') = & \sum_{i=n_x}^1 \text{LCSS}_s((x_1, \dots, x_i), x') + \\ & \sum_{i=n_{x'}-1}^1 \text{LCSS}_s(x, (x'_1, \dots, x'_i)), \end{aligned} \quad (11)$$

where

$$\text{LCSS}_s(x, x') = \begin{cases} 0, & \text{if } n_x = 0 \vee n_{x'} = 0, \\ \kappa_f(x_{n_x}, x'_{n_{x'}})(1 - v(x_{n_x}, x'_{n_{x'}})) + \text{LCSS}_s(\text{Head}(x), \text{Head}(x')), & \text{if } \kappa_f(x_{n_x}, x'_{n_{x'}}) \geq \theta, \\ \max(\text{LCSS}_s(\text{Head}(x), x'), \text{LCSS}_s(x, \text{Head}(x'))) & \text{otherwise,} \end{cases} \quad (12)$$

with $v(x_{n_x}, x'_{n_{x'}}) = 1$ if the element has been visited and 0 otherwise.

Following the definition of the all-subsequences kernel, each common subsequence s should be weighted by $2^{(n_s)}$, as all its subsets are also common subsequences. However, it turned out that weighting the matches by $2^{(n_s)}$ and then normalizing the kernel value by $2^{(\min(n_x, n_{x'}))}$ tends to favor short matching sequences, thus this weighting was discarded and normalization is done by $\min(n_x, n_{x'})$.

We have implemented the proposed kernel using the LibSVM [4] framework.

4 Evaluation

In this section we describe the experiments we have performed to validate the result from [7] that the MPEG-7 kernel performs better than the RBF kernel on the same feature vector. We then evaluate the proposed kernel for sequences of MPEG-7 features on the same data.

4.1 Data

We perform our evaluation on the TRECVID [16] 2007 High-level feature extraction data set (50 hours training data and 50 hours test data), using the ground truth from the collaborative annotation effort [1] for the training set and the truth judgments provided by NIST for the test set. These annotations are available for 20 concepts. For the sequence kernel we need more key frames than those provided in the TRECVID reference key frame set. We use a denser key frame set provided by the K-Space project [17], which has at least a few key frames for each shot and more than 100 key frames for the longest shots. All experiments are performed using this set of key frames.

We extract the following features from the key frames using our own implementations of the MPEG-7 visual feature extractors: ColorLayout, using the 3 DC coefficients, 5 AC coefficients for the Y channel and 2 AC coefficients each for the Cb/Cr channels (12 element vector), DominantColor with up to 3 dominant colors (12 element vector), ColorStructure (32 bin histogram) and EdgeHistogram (80 bin histogram).

4.2 Results

Following the definition of the TRECVID HLFE task, we use the mean average precision (MAP) of up to 2000 results (on shot granularity) per concept for evaluation.

RBF vs. MPEG-7 Kernel. For this experiment the feature vectors of the key frames of one shot are treated independently and the RBF and MPEG-7 kernels are applied to the same set of feature vectors. For the RBF kernel, the input data has been scaled to the range $[-1, 1]$. For the MPEG-7 kernel this is not necessary, as the kernels for the specific descriptors are aware of the value ranges. The RBF kernel treats the input features as a stacked feature vector and calculates the distance element-wise.

Grid search has been performed to optimize the parameters of the C-SVM (using the LibSVM [4] implementation). For both kernels the cost parameter has been selected using grid search, for the RBF kernel also the γ parameter. The MPEG-7 kernel does not have such a parameter, as the weights \bar{w}_i determined from the feature variances are used for this purpose.

Figure 1 shows the results of this experiment. For each concept, the best result obtained from the grid search procedure is given. In addition, the mean of these best runs and the best mean achieved using grid search are given. The MPEG-7 kernel outperforms the RBF kernel for 12 out of the 20 concepts, but both the

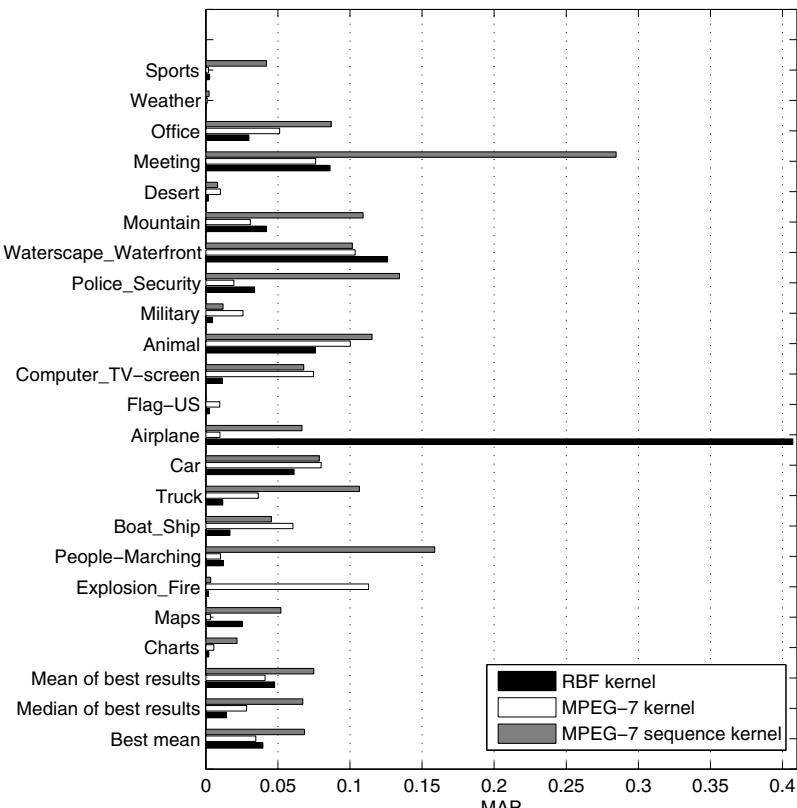


Fig. 1. MAP scores of RBF kernel, MPEG-7 kernel and MPEG-7 sequence kernel

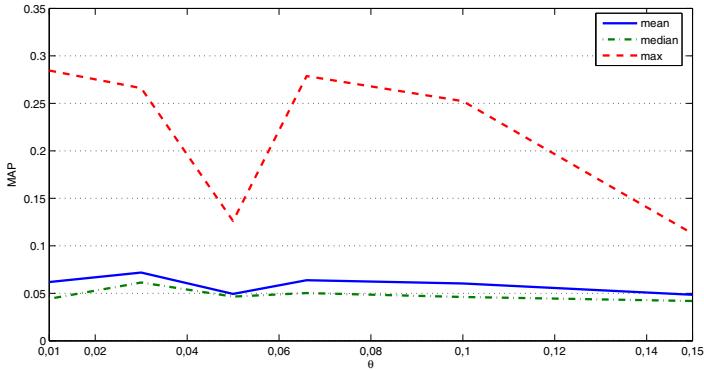


Fig. 2. Mean, median and maximum results of the MPEG-7 sequence kernel for different values of θ

mean of these runs and the best mean result are higher for the RBF kernel. However, the median MAP of the best results of the MPEG-7 kernel is nearly twice as high as that of the RBF kernel.

MPEG-7 Sequence Kernel. For the evaluation of the sequence kernel, the feature vectors of all key frames of a shot are grouped together. The kernel function then treats the elements of the sequence accordingly and calculates the LCSS based distance. The kernel has one parameter, the similarity threshold θ , which determines whether two elements are treated as similar or one element is skipped as not matching. The weights \bar{w}_i of the features are the same as in the previous experiment. No grid search has been performed, but the cost parameter has been set to 1.

Figure 2 shows the mean, median and maximum results of the concepts for different values of θ . The mean and median results show no strong dependence on the threshold value. However, the results for specific concepts may vary strongly with the threshold value as is shown by the maximum score. Also, by increasing the threshold (i.e. being more strict), the maximum score converges towards the mean.

Figure 1 compares the results to those of the RBF and MPEG-7 kernels. The MPEG-7 sequence kernel scores best for 14 of the 20 concepts and has clearly the best mean and median results of all the kernels with about 50% improvement for the mean and more than 100% improvement for the median.

Computing the weights \bar{w}_i is expensive due to the required evaluation of the distance function. If all distance functions were linear, one could calculate the variance of the input features instead of the variance of the distances which is computationally less expensive. We evaluated the results for $\theta = 0.03$. With the weights calculated from the feature variances the mean MAP decreases from 0.0684 to 0.0589 and the median MAP decreases from 0.0679 to 0.0447. As the weight calculation has only to be done once for the training set (independent of

SVM parameters) it seems worth calculating the weights using the variance of the feature distances.

4.3 Discussion

The evaluation with 7 classes and 700 images in [6] shows an improvement of precision by 0.1 when using the kernel modeling the distance functions for the MPEG-7 visual features. This result is partly confirmed by our experiment. The MPEG-7 kernel performs better for slightly more than half of the concepts (for some of them significantly better), but worse for the others. The mean MAPs seem to indicate that the RBF kernel performs overall better. However, the mean result of the RBF kernel seems to be strongly biased by the result for “Airplane”. This assumption is supported by the median MAP, which shows a clearly better performance of the MPEG-7 kernel.

The results for the MPEG-7 sequence kernel confirm results for other kernels taking temporal information into account: It clearly outperforms the two kernels just working on single key frames. As one would expect, there are strong performance improvements for dynamic concepts such as “People marching” and “Sports”. Interestingly also some rather static concepts perform better, probably due to the fact that the sequence information helps to discriminate static or low motion content from other concepts that are similar in terms of the visual features.

5 Conclusion and Future Work

In this paper we have reviewed a kernel for matching MPEG-7 visual features according to the distance functions recommended by the standard and validated the improved performance over a RBF kernel on a TRECVID data set for video concept detection. We have proposed a kernel for matching feature sequences based on the LCSS distance measure and integrating the MPEG-7 kernel. The results show that the kernel using sequences of features outperforms the kernels using only single key frame features by a significant margin. It has to be noted that grid search has not been performed for the sequence kernel, so the results can be probably improved further.

One issue with the proposed sequence kernel is the runtime due to the sequence alignment. In the experiments an existing implementation of the LCSS based distance measure has been used, and optimization options still need to be explored.

Acknowledgments

The author would like to thank Roland Mörzinger, Georg Thallinger and Werner Haas for their feedback and support.

The research leading to this paper has been partially supported by the European Commission under the contract FP7-248138, “FascinatE – Format-Agnostic SCript-based INterAcTive Experience” (<http://www.fascinate-project.eu/>).

References

1. Ayache, S., Quénnot, G.: TRECVID 2007: Collaborative annotation using active learning. In: TRECVID (2007)
2. Bailer, W., Lee, F., Thallinger, G.: A distance measure for repeated takes of one scene. *The Visual Computer* 25(1), 53–68 (2009)
3. Ballan, L., Bertini, M., Del Bimbo, A., Serra, G.: Video event classification using string kernels. *Multimedia Tools Appl.* 48(1), 69–87 (2010)
4. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines (2001), Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
5. Choi, J., Jeon, W.J., Lee, S.-C.: Spatio-temporal pyramid matching for sports videos. In: Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval, MIR 2008, pp. 291–297. ACM, New York (2008)
6. Djordjevic, D., Izquierdo, E.: Kernels in structured multi-feature spaces for image retrieval. *Electronics Letters* 42(15), 856–857 (2006)
7. Djordjevic, D., Izquierdo, E.: Relevance feedback for image retrieval in structured multi-feature spaces. In: Proceedings of the 2nd International Conference on Mobile Multimedia Communications, MobiMedia 2006, pp. 1–5. ACM, New York (2006)
8. Grauman, K., Darrell, T.: The pyramid match kernel: Efficient learning with sets of features. *J. Mach. Learn. Res.* 8, 725–760 (2007)
9. Information technology-multimedia content description interface: Part 3: Visual. ISO/IEC 15938-3 (2001)
10. Kotsia, I., Patras, I.: Relative margin support tensor machines for gait and action recognition. In: Proceedings of the ACM International Conference on Image and Video Retrieval, CIVR 2010, pp. 446–453. ACM, New York (2010)
11. Kraaij, W., Awad, G.: TRECVID-2009 high-level feature task: Overview (2009), <http://www-nplir.nist.gov/projects/tvpubs/tv9.slides/tv9.hlf.slides.pdf>
12. Manjunath, B.S., Ohm, J.-R., Vasudevan, V.V., Yamada, A.: Color and texture descriptors. *IEEE Transactions on Circuits and Systems for Video Technology* 11(6), 703–715 (2001)
13. Naphade, M.R., Kennedy, L., Kender, J.R., Chang, S.-F., Smith, J.R., Over, P., Hauptmann, A.: A light scale concept ontology for multimedia understanding for TRECVID 2005. Technical Report RC23612 (W0505-104), IBM Research (2005)
14. Qi, G.-J., Hua, X.-S., Rui, Y., Tang, J., Mei, T., Wang, M., Zhang, H.-J.: Correlative multilabel video annotation with temporal kernels. *ACM Trans. Multimedia Comput. Commun. Appl.* 5(1), 1–27 (2008)
15. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
16. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and TRECVID. In: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, MIR 2006, pp. 321–330. ACM Press, New York (2006)
17. Wilkins, P., Adamek, T., Byrne, D., Jones, G.J.F., Lee, H., Keenan, G., McGuinness, K., Smeaton, A.F., O'Connor, N.E., Amin, A., Obrenovic, Z., Benmokhtar, R., Galmar, E., Huet, B., Essid, S., Landais, R., Vallet, F., Papadopoulos, G.T., Vrochidis, S., Mezaris, V., Kompatsiaris, I., Spyrou, E., Avrithis, Y., Mörzinger, R., Schallauer, P., Bailer, W., Piatrik, T., Chandramouli, K., Izquierdo, E., Haller, M., Goldmann, L., Samour, A., Cobet, A., Sikora, T., Praks, P.: K-Space at TRECVID 2007. In: Proceedings of the TRECVID Workshop (2007)

18. Wu, G., Wu, Y., Jiao, L., Wang, Y.-F., Chang, E.Y.: Multi-camera spatio-temporal fusion and biased sequence-data learning for security surveillance. In: Proceedings of the eleventh ACM International Conference on Multimedia, MULTIMEDIA 2003, pp. 528–538. ACM, New York (2003)
19. Xu, D., Chang, S.-F.: Video event recognition using kernel methods with multilevel temporal alignment. *IEEE Trans. Pattern Anal. Mach. Intell.* 30(11), 1985–1997 (2008)
20. Yeh, M.-C., Cheng, K.-T.: A string matching approach for visual retrieval and classification. In: Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval, MIR 2008, pp. 52–58. ACM, New York (2008)

Bottom-Up Saliency Detection Model Based on Amplitude Spectrum

Yuming Fang¹, Weisi Lin¹, Bu-Sung Lee¹, Chiew Tong Lau¹, and Chia-Wen Lin²

¹ School of Computer Engineering,
Nanyang Technological University, Singapore 639798, Singapore
`{fa0001ng,wslin,ebslee,asctlau}@ntu.edu.sg`
² Department of Electrical Engineering,
National Tsing Hua University, Hsinchu, Taiwan 30013, R.O.C.
`cwlin@ee.nthu.edu.tw`

Abstract. In this paper, we propose a saliency detection model based on amplitude spectrum. The proposed model first divides the input image into small patches, and then uses the amplitude spectrum of the Quaternion Fourier Transform (QFT) to represent the color, intensity and orientation distributions for each patch. The saliency for each patch is determined by two factors: the difference between amplitude spectrums of the patch and its neighbor patches and the Euclidian distance of the associated patches. The novel saliency measure for image patches by using amplitude spectrum of QFT proves promising, as the experiment results show that this saliency detection model performs better than the relevant existing models.

Keywords: Quaternion Fourier Transform, Amplitude Spectrum, Visual Attention, Saliency Detection.

1 Introduction

When observers look at an image, they will focus on the areas they are most interested in, which are known as Regions of Interest (ROI), and this is because the human visual attention mechanism can filter the visual information to select the most important visual information and determine the eye fixation areas. According to Feature-Integration theory [6], some locations in a scene automatically stand out due to specific low-level features (such as color, intensity, orientation and so on) when observers look at this visual scene. Based on this theory, many computational models of visual attention have been advanced [3, 4, 7, 8, 9, 10, 13, 14, 16, 18, 20]. Itti L. et al. devised a visual attention model by computing the differences of color, intensity and orientation between center and surround [7]. Based on the model in [7], Harel J. et al. proposed a Graph-Based saliency detection model by using a better dissimilarity to measure the saliency [20], while Bruce N. D. et al. decided visual attention based on the information maximization [8]. Gao D. et al. calculated the center-surround discriminant for saliency detection [16]. Meur O. L. designed a bottom-up visual attention model based on the understanding of the Human Visual System (HVS) behavior [13]. In [3, 4], Hou X. et al. used phase spectrum to obtain the saliency map. Recently, a saliency detection model by Gopalakrishnan V. et al. is based on the color and orientation distributions in images [18].

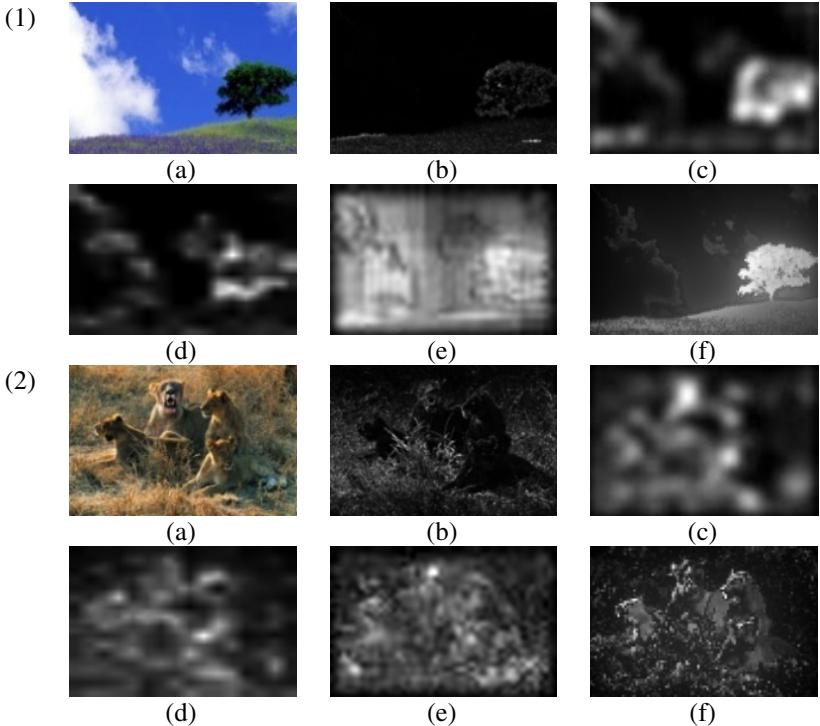


Fig. 1. The original images and the saliency maps; (a) Original images; (b) The saliency maps obtained from phase-based model without resizing the original image before FT; (c) The saliency maps obtained from phase-based model with resizing the original image into the size of 64*64 [3]; (d) The saliency maps obtained from the orientation saliency model in [18]; (e) The saliency maps obtained from [7]; (f) The saliency maps obtained from our proposed saliency detection model

It is widely accepted that the phase spectrum carries location information, while the amplitude spectrum includes the appearance and orientation information for visual scenes [1, 2, 19]. Based on this understanding, Fourier Transform (FT) has been used in computational models of human visual perception and understanding [3, 4, 5, 18]. In [3], the saliency map is actually obtained based on phase spectrum [4], while in the study [18], the amplitude spectrum of small image patches is applied to obtain the orientation saliency map. However, one problem with the phase-based models is that they have to resize the images to a smaller size to get good performance, as shown with the comparison in Fig. 1 (b) and (c). The reason is that the phase-based saliency map mainly considers the high-frequency areas in images. Although the salient object can be detected after resizing images, the detected object is different from the original, as shown in the Fig. 1 (a) and (c). The different is due to the loss of some image information caused by image resizing. Furthermore, since the phase-based model generally regards the high-frequency areas as saliency, it ignores the salient objects made up by low-frequency areas, as shown in Fig. 1 (2) (b) and (c), where the saliency maps fail to include the tigers which contain the low-frequency areas. From Fig. 1 (d),

we can see that the orientation saliency detection model in [18] based upon amplitude spectrum also suffers from the problem that the shape of the salient objects is different from the original. This is because the orientation saliency map of [18] is calculated by the histogram of some special orientations for patches, which causes the loss of some orientation information. The classical saliency detection model [7] uses the local color, intensity and orientation contrast between the center and the surrounding to obtain the saliency map. One problem with this model is that it might regard the non-salient areas with high local contrast as saliency, as shown in Fig. 1 (e). In the saliency map from Fig. 1 (2) (e), some background areas with high local contrast are considered as salient areas. In the proposed model we achieve much better performance by using all the amplitude information of image patches to get the saliency map, as shown in Fig. 1(f). We will explain the proposed model in detail and show more experiment results in the following sections.

In this paper, we propose a new saliency detection model based on the amplitude spectrum to achieve a higher accuracy in detection of salient areas. According to the Feature-Integration theory [6], we know that the patches which have a greater difference from their neighbors are considered more salient. The salient value for an image patch can be measured by the differences between its amplitude spectrum and its neighbors', as well as the Euclidian distance between this patch and its neighbors. Here we use the QFT to obtain the amplitude spectrum to represent image patches. Compared with [18] in which the color and orientation distributions for image patches are obtained in two different methods, the proposed model is much simpler because it gets all the color, intensity and orientation distributions by using amplitude spectrum of QFT. In addition, in [18], the orientation histogram is used to compute the differences between image patches, which will cause the loss of orientation information. In this paper, we use more accurate algorithms to calculate the differences between image patches to measure the saliency of each patch. Therefore, the saliency value of image patches in the proposed model will be more accurate, and this is demonstrated in Section 3.

In Section 2, we will describe the details of this proposed model. Section 3 reports on the experiment results carried out for comparing our new model with relevant existing methods. Finally, we conclude the paper by summarizing our findings in Section 4.

2 Approach

As described in the preceding section, the input image is first divided into small image patches. Each image patch is represented by its intensity and color. By using Quaternion Fourier Transform (QFT) [11] on each patch, we obtain the amplitude spectrum of each patch. Computing the Euclidian distance between an image patch and its neighbor patches, and the differences between the amplitude spectrum of this image patch and its neighbor patches, we get the saliency value for this patch. Finally, we obtain the saliency map from the saliency value of each patch. The architecture of the proposed model is shown in Fig. 2.

According to the Feature-Integration theory, we measure the saliency of each image patch by calculating the differences between this patch and its neighbors. If the differences between a patch and its neighbors are larger, then the saliency value of this patch is larger. Similar to the orientation saliency map in [18], we add the influence of spatial information in the proposed model. In the spatial domain, with the distance

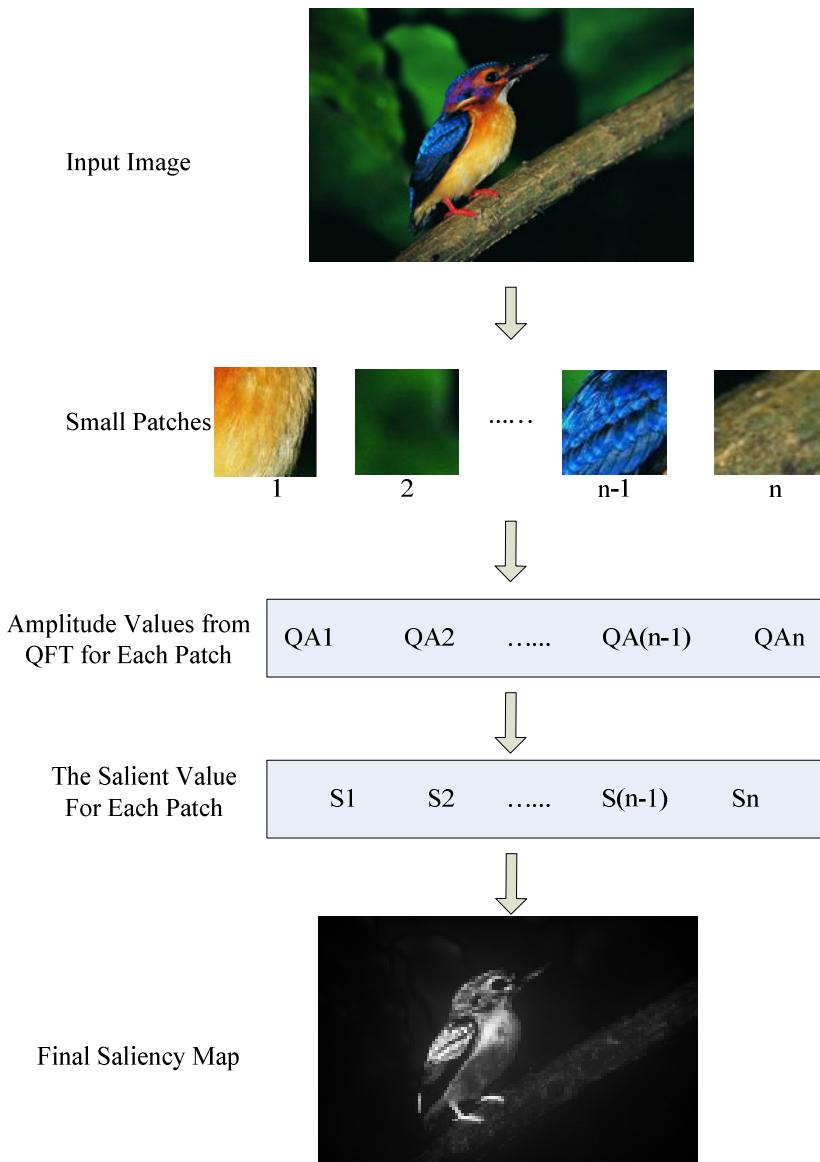


Fig. 2. The architecture of the proposed model

between the patch and its neighbor increasing, the contribution of the difference to the saliency value for this patch is reducing. Here we use $Df(i,j)$ to represent the difference between patch i and patch j. The salient value S_i for image patch i can be calculated as follows:

$$S_i = \sum_{i \neq j} \frac{Df(i,j)}{D(i,j)} \quad (1)$$

where $D(i, j)$ is the Euclidian distance between patch i and patch j .

From Equation (1), the salient value of each patch is obtained from the sum of differences between this patch and all its neighbors. These neighbors include all the patches in the input image except itself. The contribution from the difference of more distant neighbor to the patch's salient value is smaller. We will further discuss this and its improvement over the relevant existing models in the following subsections.

2.1 Obtaining the Amplitude Values for Each Patch

Here we use the color and intensity channels for QFT to get the amplitude spectrum for each image patch, which are used to compute the difference between two image patches. Compared with FT used in [18], QFT can be used to describe the spectral content of color images [11]. In this model, each image patch is represented by two color channels and one intensity channel. For each image patch, r , g and b denote the red, green and blue color components respectively. The intensity channel can be computed as [7]:

$$I = (r + g + b)/3 \quad (2)$$

We can get four broadly-tuned color channels as follows:

$$R = r - (g + b)/2 \quad (3)$$

$$G = g - (r + b)/2 \quad (4)$$

$$B = b - (r + g)/2 \quad (5)$$

$$Y = \frac{r+g}{2} - \frac{|r-g|}{2} - b \quad (6)$$

Finally we arrive at two new color channels of red/green and blue/yellow to be used in the model as [4, 12]:

$$RG = R - G \quad (7)$$

$$BY = B - Y \quad (8)$$

The QFT [11] result of I , RG and BY for an image patch is given as:

$$QR = \mathcal{F}(I, RG, BY) \quad (9)$$

where I , RG and BY are intensity and color channels, and \mathcal{F} is the QFT.

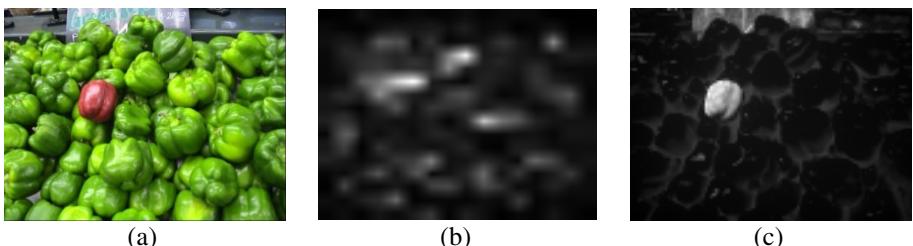


Fig. 3. The comparison between the saliency maps generated from [18] and the proposed model: (a) the original image; (b) the orientation saliency map from [18]; (c) the saliency map from the proposed model

We can write QR as:

$$QR = A e^{i\varphi} \quad (10)$$

where A is the amplitude of the image patch and φ is the corresponding phase.

From Equation (10), we can get the amplitude spectrum for an image patch. In this model, we use this amplitude spectrum to represent each image patch. As we use the color and intensity channels for QFT, this amplitude spectrum includes color information as well as intensity information. As we know, the amplitude spectrum indicates the presence of the respective spatial frequencies and their strengths can represent the orientation distribution in images [18, 19]. Thus, the amplitude spectrum of QFT for an image patch can represent the color, intensity and orientation distributions for this patch. However, the model in [18] just uses FT on gray image patches to obtain the orientation distribution. Thus, the orientation saliency map in [18] does not include the color information, as shown in Fig. 3.

2.2 Salient Value for Each Patch

As described previously, the salient value for each image patch is determined by the differences between this patch and its neighbors. If a patch is very different from its neighbors, it will be a salient area. The salient value for a patch is larger when the difference between this patch and its neighbor is greater. However, with the increasing distance between the patch and its neighbors, the contribution of this difference to the salient value for the patch decreases. The salient value for an image patch can be calculated according to Equation (1).

As described above, the amplitude spectrum of QFT represents the color, intensity and orientation distributions for image patches. In the study of [18], the authors used the dissimilarity of the entropy of the orientation to represent the orientation distribution difference between the image patch and its neighbors; whereas in this paper we use two different algorithms to represent the differences between each patch and its neighbors. These differences in the proposed model cover the dissimilarities of color, intensity and orientation distributions of image patches. From the experiment results, we find that both two algorithms of the proposed model are much better than that in [18].

2.2.1 Algorithm 1

In this algorithm, the sum of amplitude spectrum for an image patch is used to represent the patch. In addition, the Euclidian distance between two patches represents the difference between these two patches. We use logarithm to reduce the dynamic range of the amplitude coefficients for better effect and add the constant 1 to each original amplitude value to avoid the undefined case when A approaches zero. Thus, the difference between image patch i and j can be calculated as follows:

$$Df(i, j) = |\sum_m \log(A_m^i + 1) - \sum_n \log(A_n^j + 1)| \quad (11)$$

where $\sum_m \log(A_m^i + 1)$ is the sum of the logarithm of amplitude values for patch i ; $\sum_n \log(A_n^j + 1)$ is the sum of the logarithm of amplitude values for patch j . The final saliency map is formed by the saliency value for each patch.

2.2.2 Algorithm 2

We also try to use another algorithm to represent the difference between two image patches: to calculate the Euclidian distance between the amplitude spectrum of two image patches as the difference between them. Using this method, we can calculate the difference between image patches i and j as follows:

$$Df(i, j) = \sqrt{\sum_m (\log(\mathcal{A}_m^i + 1) - \log(\mathcal{A}_m^j + 1))^2} \quad (12)$$

Like Algorithm 1, this algorithm includes the differences of color, intensity and orientation distributions for two image patches. Compared with Algorithm 1, the difference between two image patches in this algorithm is calculated more accurately. Experiment results in the following section show that using this algorithm can get the better saliency map compared with that of Algorithm 1.

3 Experiments

The experiment comparison includes two parts: one focuses on comparing the effect of saliency maps generated by the proposed model and others; the other is qualitative overall performance evaluation.

We choose three existing models to compare with the proposed model: Itti's model [7], Hou's model [3] and Gopalakrishnan's model [18]. The reason for our choice of these three models is: Itti's model is a classical visual model; Hou's model is a phase-based model which also uses FT; Gopalakrishnan's model is most related to the proposed model because they use the amplitude spectrum of FT to obtain the orientation saliency map. The images in this comparison are from the image database in [3] and we use some experiment results in [3]. Experiments show that there is little influence to the effect of saliency map for different patch sizes. However, the patch size is inversely proportional to the computational complexity. In the experiments, we divide images into 6×6 image patches. In Fig. 4, we compare the saliency detection maps obtained from our proposed model and from other models. This figure includes the results of four types of images from the image database in [3].

The first column of Fig. 4 shows the saliency map for an image including only a single object, from which we can see that the saliency map from Hou's model is better than those from Itti's and Gopalakrishnan's models. The proposed model outperforms Hou's model for this image. The reason is that input image is resized into the size of 64×64 in advance in Hou's model. This downsampling process causes the loss of some image information. While the proposed model preserves the original size for the input image and the salient areas generated from this model are better than those from the other models. In addition, the saliency map generated from the proposed model has the highest similarity to the human labeled map.

The second column of Figure 4 indicates the saliency map for the image including a single object and its shadow. From the human labeled map, the shadow is also the saliency area. However, results from all other three models [3, 7, 18] do not include this area. Both saliency maps from the proposed model using Algorithm 1 and Algorithm 2 can detect this salient shadow.

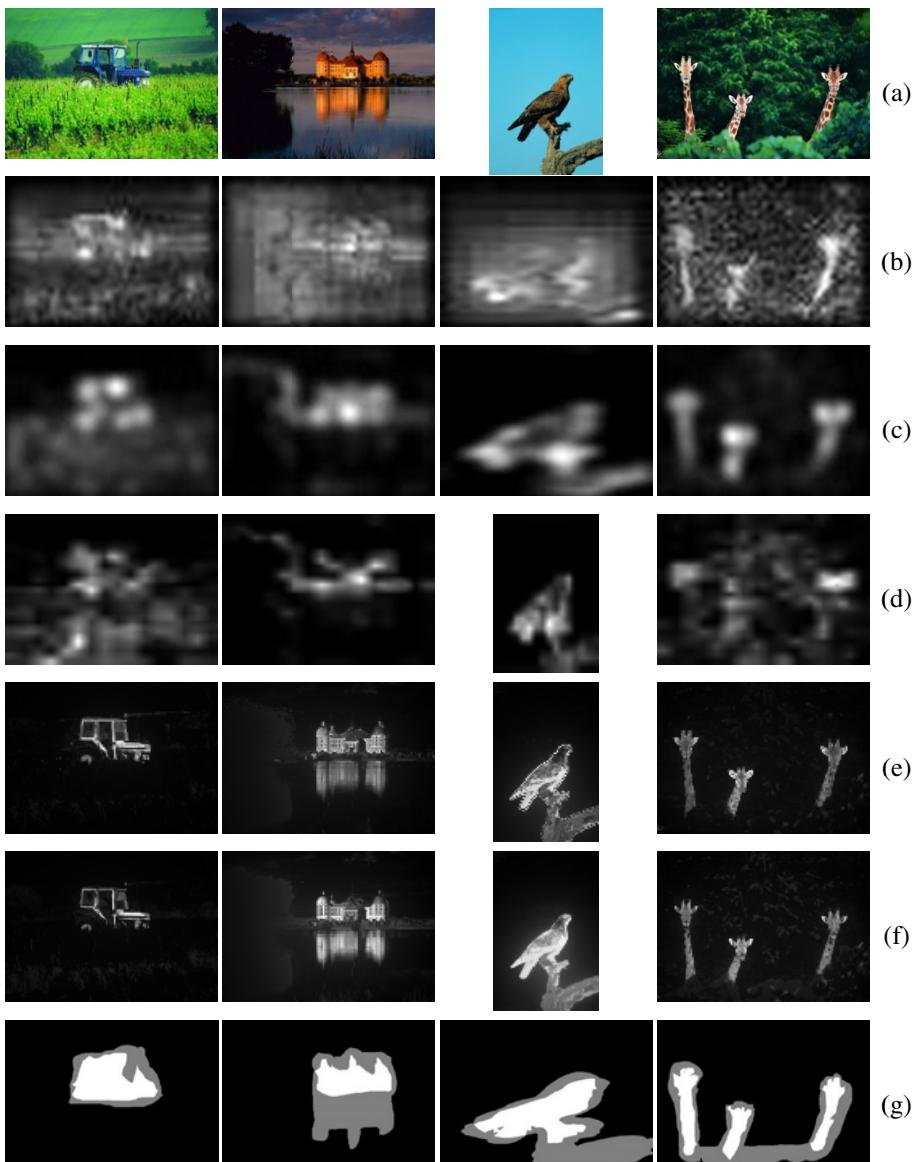


Fig. 4. Saliency maps for different saliency detection models: (a) The original images; (b) Saliency maps from Itti's model [7]; (c) Saliency maps from Hou's model [3]; (d) Saliency maps from Gopalakrishnan's model [18]; (e) Saliency maps from the proposed model using Algorithm 1; (f) Saliency maps from the proposed model using Algorithm 2; (g) Human labeled maps. The saliency maps of (b), (c) and (g) are from the original experiment results in [3].

Table 1. The ROC areas for different models with all human fixations (The proposed model is with the Algorithm 2 to represent each patch)

Saliency Detection Model	Itti's Model [15]	Bruce's Model [8]	Gao's Model [16]	The Proposed Model
ROC Area	0.7287	0.7547	0.7694	0.7917

The third column of Figure 4 shows the saliency map for the image including a bird standing on the branch. Compared with other models, the proposed model can detect the object region more clearly. From this column, we can also see that the saliency map from the proposed model using Algorithm 2 is better than that using Algorithm 1. The reason for this is that using the Algorithm 2 can obtain more detailed information from the patch compared with that of using Algorithm 1. Thus, for better effect, we use the proposed model with Algorithm 2 to perform the next experiment.

The fourth column of Figure 4 shows the saliency map from a multi-object image. Although all models can detect the three objects in this image, the object areas from the saliency map of our model are more accurate than those from others. From this column and the first column, we can see that Gopalakrishnan's model is greatly influenced by the context with strong orientation, because it uses the histogram of orientation information during calculating the dissimilarity and this causes the model to be more sensitive to the stronger orientation.

Overall, from Fig. 4, we can clearly see that the effect of saliency map from the proposed model is much better than those from other models. The salient objects are clearer and the saliency maps obtained are closer to the human labeled maps when compared with those from other models.

A saliency map is commonly used to predict human eye fixation. Thus, one way to evaluate the performance of saliency detection model is to compare the saliency map with human eye fixation. We therefore further compare the performance of the proposed model with other models [8, 15, 16] with respect to human fixation data. The image dataset and human fixation data are from [8]. The database includes 120 images with size of 681*581 and their corresponding eye tracking data. Using the evaluation metric in [17], we can get the experiment results as shown in Table 1.

In Table 1, the receiver operating characteristic (ROC) area results of other models [8, 15, 16] are extracted from the original experiment results in [16]. From the results, we can see that our model performs better than other models in overall agreement with human fixation results.

4 Discussions and Conclusions

The proposed model uses the amplitude spectrum of QFT to represent the color, intensity and orientation distributions for image patches. It measures the saliency of each patch through calculating the differences between amplitude spectrums of this patch and its whole neighbor patches in the image. The novel method of measuring saliency for each image patch has proven to be encouraging. Compared with the saliency map in [18] which uses two different frameworks to compute the color and orientation distributions respectively, we obtain the color, intensity and orientation

distributions through the amplitude spectrum of QFT. In addition, the orientation saliency map in [18] is computed by using the orientation histogram which will cause loss of some information for image patches. Thus, although the proposed saliency detection model is much simpler than that of [18], the saliency map obtained from the proposed model is much better than that of [18]. As for the phase-based saliency detection models in [3, 4], the saliency map mainly includes the high-frequency areas in images, while ignores the low-frequency salient areas. The proposed model obtains the saliency map through the differences between image patches, which consider both cases of low-frequency salient areas and high-frequency salient areas. As discussed in Section 1, the classical saliency detection model in [7] may regard the non-salient areas with local high contrast as salient and thus misleads the saliency computation. Another problem with this classical model is that the saliency map is influenced greatly by the context and thus the salient objects are not clear. In the proposed model, as we consider the influence of all other image patches when we calculate the saliency value for an image patch, the problems with the classical model of [7] do not exist.

In summary, we have proposed a new saliency detection model based on amplitude spectrum of QFT in this paper in order to overcome the aforementioned shortcomings of the relevant models in the literature. It first divides the input image into small image patches. Through using QFT to the image patches based on color and intensity channels, the model uses the amplitude spectrum to represent the color, intensity and orientation distributions for the image patches. The salient value for each patch is obtained by computing the differences between the amplitude spectrums of the patch and its neighbors. Experiment results have shown that the proposed saliency detection model can more accurately detect the saliency area. Compared with other existing models, the proposed model has better performance with regard to ground truth of the human eye fixation.

Acknowledgments. The authors would like to thank Neil D. B. Bruce and Hou X. for their image database. The authors also would like to thank Viswanath Gopalakrishnan for his source code and helpful discussions, and Steve Sangwine for his helpful advice in using the source code of QFT.

References

1. Oppenheim, A.V., Lim, J.S.: The importance of phase in signals. Proc. IEEE 69, 529–541 (1981)
2. Piotrowski, L.N., Campbell, F.W.: A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. Perception 11, 337–346 (1982)
3. Hou, X., Zhang, L.: Saliency Detection: A spectral residual approach. In: Proceedings of IEEE CVPR (2007)
4. Guo, C., Ma, Q., Zhang, L.: Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In: Proceedings of IEEE CVPR (2008)
5. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelop. International Journal of Computer Vision 42, 145–175 (2001)
6. Treisman, A., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology 12(1), 97–136 (1980)

7. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(11), 1254–1259 (1998)
8. Bruce, N.D., Tsotsos, J.K.: Saliency based on information maximization. *Advances in Neural Information Processing Systems* 18, 155–162 (2006)
9. Brecht, M.D., Saiki, J.: A neural network implementation of a saliency map model. *Neural Networks* 19, 1467–1474 (2006)
10. Lu, Z., Lin, W., Yang, X., Ong, E., Yao, S.: Modeling visual attention’s modulatory aftereffects on visual sensitivity and quality evaluation. *IEEE Transactions on Image Processing* 14(11), 1928–1942 (2005)
11. Ell, T., Sangwin, S.: Hypercomplex Fourier Transforms of Color Images. *IEEE Transactions on Image Processing* 16(1), 22–35 (2007)
12. Engel, S., Zhang, X., Wandell, B.: Colour Tuning in Human Visual Cortex Measured With Functional Magnetic Resonance Imaging. *Nature* 388(6), 68–71 (1997)
13. Le Meur, O., Le Callet, P., Barba, D., Thoreau, D.: A coherent computational approach to model bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 28(5), 802–817 (2006)
14. Liu, T., Sun, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. In: Proc. CVPR (2007)
15. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. *Visual Research* 40, 1489–1506 (2000)
16. Gao, D., Vasconcelos, N.: Bottom-up saliency is a discriminant process. In: Proceedings of IEEE CVPR (2007)
17. Tatler, B.W., Baddeley, R.J., Gilchrist, I.D.: Visual correlates of fixation selection: effects of scale and time. *Visual Search* 45, 643–659 (2005)
18. Gopalakrishnan, V., Hu, Y., Rajan, D.: Salient region detection by modeling distributions of color and orientation. *IEEE Transactions on Multimedia* 11(5), 892–905 (2009)
19. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*, 3rd edn. Prentice-Hall, Englewood Cliffs
20. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: Proc. NIPS (2006)

L_2 -Signature Quadratic Form Distance for Efficient Query Processing in Very Large Multimedia Databases

Christian Beecks, Merih Seran Uysal, and Thomas Seidl

Data Management and Data Exploration Group

RWTH Aachen University, Germany

{beecks, uysal, seidl}@cs.rwth-aachen.de

Abstract. The highly increasing amount of multimedia data leads to extremely growing databases which support users in searching and exploring the database contents. Content-based searching for similar objects inside such vivid and voluminous multimedia databases is typically accompanied by an immense amount of costly similarity computations among the stored data objects. In order to process similarity computations arising in content-based similarity queries efficiently, we present the L_2 -Signature Quadratic Form Distance which maintains high retrieval quality and improves the computation time of the Signature Quadratic Form Distance by more than one order of magnitude. As a result, we process millions of similarity computations in less than a few seconds.

Keywords: Signature Quadratic Form Distance, feature signature, content-based retrieval, multimedia database, database searching, database query processing.

1 Introduction

Over the last couple of years, the amount of multimedia data has highly increased in volume, encouraged by the rapid spread of cheap and user-friendly multimedia devices. These devices enable users to readily generate hundreds of multimedia objects including images, videos, or music which are then processed and finally stored in voluminous multimedia databases. In this way, billions of multimedia objects become publicly available and thus numerous multimedia repositories emerge which aim at supporting users and systems in searching and exploring their quickly growing contents.

Content-based similarity queries, which frequently arise in multimedia databases for the purpose of content searching and exploring, are typically accompanied by an immense amount of costly similarity computations among the stored data objects. In order to determine the most relevant objects which are finally returned to the user, the objects' inherent properties are compactly stored in some kind of content representation and then compared by means of similarity measures. As the retrieval performance in terms of efficiency and effectiveness

depends on the applied similarity measure, we focus on the Signature Quadratic Form Distance [13] showing good applicability to different types of media and high quality in comparing flexible content representations with each other [2].

In the present paper, we propose the *L₂-Signature Quadratic Form Distance* for efficient query processing in very large multimedia databases. The L₂-Signature Quadratic Form Distance maintains high retrieval quality and effectively improves the computation time of the conventional Signature Quadratic Form Distance by more than one order of magnitude. By improving the efficiency of content-based similarity queries, we are able to process millions of similarity computations in less than a few seconds.

The structure of this paper is as follows: we first present the applied similarity model, namely the Signature Quadratic Form Distance on feature signatures, in Section 2. Then, we introduce the L₂-Signature Quadratic Form Distance in Section 3. We evaluate our approach on different image databases in Section 4 and conclude this paper with a brief outlook on future work in Section 5.

2 Content Representation Forms and Similarity Measures of Multimedia Data

In this section, we present preliminary information involving the applied content representation forms and similarity measures of multimedia data.

In order to digitize and compactly store multimedia objects' inherent properties in form of a content representation, we first need to define an appropriate feature space capturing these properties. For instance in the context of images, the feature space may comprise color, texture, position, or other local information arising from the extraction of local color descriptors [4][2][17][19]. The content of each multimedia object is then exhibited via its feature distribution in the feature space, i.e. in the context of images each image pixel is mapped to a single feature in this space. In other words, each image is represented by a set of features. We illustrate this step, *feature extraction*, in Figure 1 where the database objects and their features are depicted via the same color.

To efficiently compute similarity between two multimedia objects and thus their feature distributions by means of adaptive similarity measures, local features are aggregated into a compact content representation form. There are two common representation forms which can be compared by adaptive similarity measures, namely *feature histograms* and *feature signatures*, which arise from global partitioning of the feature space and individual clustering of features for each multimedia object, respectively [6]. Both different content representation forms are the outcome of the second step, *feature aggregation*, which is illustrated on the right-hand side of Figure 1. The global partitioning of the feature space, indicated by $\mathcal{P}_1, \dots, \mathcal{P}_4$, is carried out regardless of feature distributions of individual objects. For each object, a single feature histogram is generated according to the global partitioning where each entry of the feature histogram corresponds to the number of features located in the corresponding global partition. In contrast, individual clustering of features for a multimedia object results in a feature

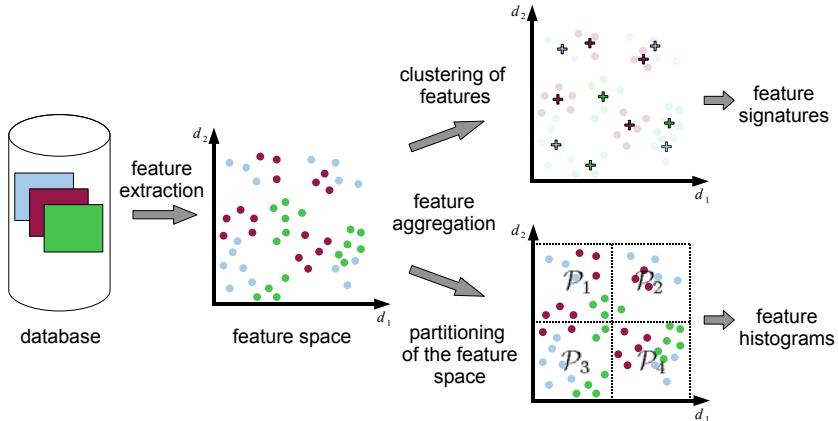


Fig. 1. The feature extraction process in two steps: the objects' properties are extracted and mapped in some feature space and then represented in a more compact form [2]

signature, also named as *adaptive-binning histogram* [10], which comprises clusters of the object's features with the corresponding weights and centroids. As can be seen in the figure, the resulting feature signature consists of clusters of the object's features with the corresponding weights and centroids, visualized by plus signs with different colors for each data object. The resulting feature signatures of the red and blue objects contain four centroids while the green object's feature signature comprises three centroids. In this way, each multimedia object is represented by an individual feature signature which is formally defined below.

Definition 1. Feature Signature

Given a feature space \mathcal{FS} and a clustering $\mathcal{C} = \mathcal{C}_1, \dots, \mathcal{C}_n$ of features $f_1, \dots, f_k \in \mathcal{FS}$ of object o , the feature signature S^o is defined as a set of tuples from $\mathcal{FS} \times \mathcal{R}^+$ as follows:

$$S^o := \{\langle c_i^o, w_i^o \rangle, i = 1, \dots, n\},$$

where $c_i^o = \frac{\sum_{f \in \mathcal{C}_i} f}{|\mathcal{C}_i|}$ and $w_i^o = \frac{|\mathcal{C}_i|}{k}$ represent the centroid and weight, respectively.

Intuitively, a feature signature S^o of object o is a set of centroids $c_i^o \in \mathcal{FS}$ with the corresponding weights $w_i^o \in \mathcal{R}^+$ of the clusters \mathcal{C}_i . The weights w_i^o of each feature signature S^o are normalized, i.e. it holds that $\sum_{i=1}^n w_i^o = 1$. Carrying out the feature clustering individually for each object, we recognize that each feature signature reflects the feature distribution more meaningfully than a feature histogram. In addition, feature histograms can be regarded as a special case of feature signatures where the clustering of object features is replaced with a global partitioning of the feature space. By assigning a centroid to each partition, feature histograms can thus be stored in form of feature signatures.

In Figure 2, we depict three example images in the top row with the visualization of the corresponding feature signatures in the bottom row. We visualize the feature signatures' centroids from a seven-dimensional feature space (two position, three color, and two texture dimensions) as circles in a two-dimensional

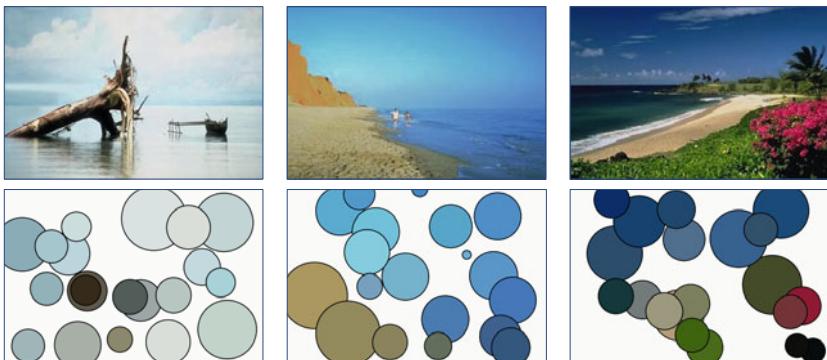


Fig. 2. Three example images in the top row and the visualization of their feature signatures in the bottom row

position space. The circles' colors and diameters reflect the colors and weights of the centroids, respectively. It can be recognized in the figure that the feature signatures reflect the visual contents of the corresponding images in an appropriate way and that the feature signatures' centroids adjust to individual images by only storing the information exhibited in the images' feature distributions. In this example only the feature signature of the image on the right-hand side consists of centroids reflecting reddish and greenish colors, while all other feature signatures comprise centroids reflecting bluish and brownish colors. To sum up, feature signatures achieve a better balance between expressiveness and efficiency than feature histograms and they can adjust to the individual images more flexible than feature histograms [15].

So far, we have presented two different forms of content representation, namely feature signatures and feature histograms. In the following of this section, we will proceed with adaptive similarity measures. To compare the contents of multimedia objects, distances frequently serve as similarity measures among the stored content representations, such as feature histograms or feature signatures. The lower the distance between two feature histograms or feature signatures, the higher the similarity between the corresponding multimedia objects and vice versa. Distances for feature histograms can be found in [7]. However, most of them are only applicable to feature histograms sharing a global partitioning and thus exhibiting the same structure and size.

As our focus lies on comparing feature signatures by making use of adaptive similarity measures we will continue with presenting the *Signature Quadratic Form Distance* [13] which is able to compare feature histograms and feature signatures of different size and structure. It has been shown, that the Signature Quadratic Form Distance produces good retrieval results and outperforms the major state-of-the-art similarity measures including the *Hausdorff Distance* [9], *Perceptually Modified Hausdorff Distance* [14], *Weighted Correlation Distance* [10], and *Earth Mover's Distance* [16] in terms of efficiency and effectiveness [2]. In the following, we give the definition of the Signature Quadratic Form Distance.

Definition 2. *Signature Quadratic Form Distance*

Given two feature signatures $S^q = \{\langle c_i^q, w_i^q \rangle | i=1, \dots, n\}$ and $S^p = \{\langle c_i^p, w_i^p \rangle | i=1, \dots, m\}$, and a similarity function $f_s(c_i, c_j) \mapsto \mathcal{R}$, the Signature Quadratic Form Distance $SQFD_{f_s}$ between S^q and S^p is defined as:

$$SQFD_{f_s}(S^q, S^p) := \sqrt{(w^q - w^p) \cdot A_{f_s} \cdot (w^q - w^p)^T},$$

where $A_{f_s} \in \mathcal{R}^{(n+m) \times (n+m)}$ is the similarity matrix arising from applying the similarity function f_s to the corresponding centroids, i.e. $a_{ij} = f_s(c_i, c_j)$. Furthermore, $w^q = (w_1^q, \dots, w_n^q)$ and $w^p = (w_1^p, \dots, w_m^p)$ form weight vectors, and $(w^q - w^p) = (w_1^q, \dots, w_n^q, -w_1^p, \dots, -w_m^p)$ denotes the concatenation of w^q and $-w^p$.

According to Definition 2, the Signature Quadratic Form Distance computes the similarity matrix A by making use of a similarity function [13] such as $f_s(c_i, c_j) = e^{-L_2^2(c_i, c_j)/2}$ modeling the similarity between two centroids c_i and c_j . In this way, the similarity matrix A is dynamically determined for each comparison of two feature signatures and it depends on the order of the feature signatures' centroids. It reflects similarities among all centroids of both feature signatures. This similarity measure enables the comparison of feature signatures and also feature histograms of any lengths regarding the quadratic form distance concept and it is a generalization of the Quadratic Form Distance [518].

Based on the presented forms of content representation and the Signature Quadratic Form Distance, we introduce a novel, specific instance of the latter in the next section.

3 L₂-Signature Quadratic Form Distance

In this section, we investigate the Signature Quadratic Form Distance with the aim of reducing its computational complexity in order to process similarity queries in very large multimedia databases efficiently.

Although the Signature Quadratic Form Distance exhibits a low default computational complexity compared with the other state-of-the-art similarity measures [1,2,3], such as the Earth Mover's Distance, we state that it is infeasible to process millions of distance computations among feature signatures, arising in content-based similarity queries, in less than a few seconds. In order to overcome this computational gap and improve the distance's efficiency, we first define the L₂-Signature Quadratic Form Distance, and then show its equivalence to the Euclidean distance based on the weighted means of the feature signatures. The L₂-Signature Quadratic Form Distance is defined as below.

Definition 3. *L₂-Signature Quadratic Form Distance*

Given two feature signatures $S^q = \{\langle c_i^q, w_i^q \rangle | i=1, \dots, n\}$ and $S^p = \{\langle c_i^p, w_i^p \rangle | i=1, \dots, m\}$ and the similarity function $f_{L_2}(c_i, c_j) = -\frac{1}{2} \cdot L_2^2(c_i, c_j)$ where L_2 denotes the Euclidean distance, the L₂-Signature Quadratic Form Distance L₂-SQFD between S^q and S^p is defined as:

$$L_2-SQFD(S^q, S^p) := SQFD_{f_{L_2}}(S^q, S^p).$$

According to Definition 3, the L_2 -Signature Quadratic Form Distance is a specific instance of the Signature Quadratic Form Distance making use of the *squared Euclidean similarity function* $f_{L_2}(c_i, c_j) = -\frac{1}{2} \cdot L_2^2(c_i, c_j)$ for computing each similarity matrix entry a_{ij} between the corresponding centroids c_i and c_j . Based on this similarity function, we will show in the following theorem how to simplify the L_2 -Signature Quadratic Form Distance computation.

Theorem 1. Efficient Computation of L_2 -SQFD

Given two feature signatures $S^q = \{\langle c_i^q, w_i^q \rangle, i = 1, \dots, n\}$ and $S^p = \{\langle c_i^p, w_i^p \rangle, i = 1, \dots, m\}$, for the L_2 -Signature Quadratic Form Distance L_2 -SQFD between S^q and S^p it holds that:

$$L_2\text{-SQFD}(S^q, S^p) = L_2(\bar{c}^q, \bar{c}^p),$$

where $\bar{c}^q = \frac{\sum_{i=1}^n w_i^q \cdot c_i^q}{\sum_{i=1}^n w_i^q}$ and $\bar{c}^p = \frac{\sum_{i=1}^m w_i^p \cdot c_i^p}{\sum_{i=1}^m w_i^p}$ are the weighted mean centroids of feature signatures S^q and S^p , respectively, and L_2 denotes the Euclidean distance between these centroids.

Proof. For better legibility, we show the statement above for the squared L_2 -Signature Quadratic Form Distance. We denote the dimensionality of the feature space \mathcal{FS} by the parameter $\dim \in \mathcal{N}^+$.

$$\begin{aligned} L_2\text{-SQFD}^2(S^q, S^p) &= (w^q| - w^p) \cdot A_{f_{L_2}} \cdot (w^q| - w^p)^T \\ &= \sum_{i=1}^n \sum_{j=1}^n w_i^q w_j^q \cdot a_{ij} + \sum_{i=1}^m \sum_{j=1}^m w_i^p w_j^p \cdot a_{ij} - 2 \sum_{i=1}^n \sum_{j=1}^m w_i^q w_j^p \cdot a_{ij} \\ &= - \sum_{i=1}^n \sum_{j=1}^n w_i^q w_j^q \frac{L_2^2(c_i^q, c_j^q)}{2} - \sum_{i=1}^m \sum_{j=1}^m w_i^p w_j^p \frac{L_2^2(c_i^p, c_j^p)}{2} \\ &\quad + 2 \sum_{i=1}^n \sum_{j=1}^m w_i^q w_j^p \frac{L_2^2(c_i^q, c_j^p)}{2} \\ &= -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_i^q w_j^q \cdot \sum_{k=1}^{\dim} (c_{i,k}^q - c_{j,k}^q)^2 - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m w_i^p w_j^p \cdot \sum_{k=1}^{\dim} (c_{i,k}^p - c_{j,k}^p)^2 \\ &\quad + \sum_{i=1}^n \sum_{j=1}^m w_i^q w_j^p \cdot \sum_{k=1}^{\dim} (c_{i,k}^q - c_{j,k}^p)^2 \\ &= -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_i^q w_j^q \cdot \sum_{k=1}^{\dim} c_{i,k}^q {}^2 + \sum_{i=1}^n \sum_{j=1}^n w_i^q w_j^q \cdot \sum_{k=1}^{\dim} (c_{i,k}^q \cdot c_{j,k}^q) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_i^q w_j^q \cdot \sum_{k=1}^{\dim} c_{j,k}^q {}^2 \\ &\quad - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m w_i^p w_j^p \cdot \sum_{k=1}^{\dim} c_{i,k}^p {}^2 + \sum_{i=1}^m \sum_{j=1}^m w_i^p w_j^p \cdot \sum_{k=1}^{\dim} (c_{i,k}^p \cdot c_{j,k}^p) - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m w_i^p w_j^p \cdot \sum_{k=1}^{\dim} c_{j,k}^p {}^2 \\ &\quad + \sum_{i=1}^n \sum_{j=1}^m w_i^q w_j^p \cdot \sum_{k=1}^{\dim} c_{i,k}^q {}^2 - 2 \sum_{i=1}^n \sum_{j=1}^m w_i^q w_j^p \cdot \sum_{k=1}^{\dim} (c_{i,k}^q \cdot c_{j,k}^p) + \sum_{i=1}^n \sum_{j=1}^m w_i^q w_j^p \cdot \sum_{k=1}^{\dim} c_{j,k}^p {}^2 \end{aligned}$$

$$\begin{aligned}
 &= -\frac{1}{2} \underbrace{\sum_{i=1}^n \sum_{j=1}^n w_i^q w_j^q \cdot \sum_{k=1}^{\dim} c_{i,k}^q}_{=0}^2 - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_i^q w_j^q \cdot \sum_{k=1}^{\dim} c_{j,k}^q \cdot \sum_{k=1}^{\dim} c_{j,k}^q + \sum_{i=1}^n \sum_{j=1}^m w_i^q w_j^p \cdot \sum_{k=1}^{\dim} c_{i,k}^q \cdot \sum_{k=1}^{\dim} c_{i,k}^q \\
 &\quad - \frac{1}{2} \underbrace{\sum_{i=1}^m \sum_{j=1}^m w_i^p w_j^p \cdot \sum_{k=1}^{\dim} c_{i,k}^p}_{=0}^2 - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m w_i^p w_j^p \cdot \sum_{k=1}^{\dim} c_{j,k}^p \cdot \sum_{k=1}^{\dim} c_{j,k}^p + \sum_{i=1}^n \sum_{j=1}^m w_i^q w_j^p \cdot \sum_{k=1}^{\dim} c_{j,k}^p \\
 &\quad + \sum_{i=1}^n \sum_{j=1}^n w_i^q w_j^q \cdot \sum_{k=1}^{\dim} (c_{i,k}^q \cdot c_{j,k}^q) + \sum_{i=1}^m \sum_{j=1}^m w_i^p w_j^p \cdot \sum_{k=1}^{\dim} (c_{i,k}^p \cdot c_{j,k}^p) \\
 &\quad - 2 \sum_{i=1}^n \sum_{j=1}^m w_i^q w_j^p \cdot \sum_{k=1}^{\dim} (c_{i,k}^q \cdot c_{j,k}^p) \\
 &= \sum_{k=1}^{\dim} \sum_{i=1}^n (c_{i,k}^q \cdot w_i^q) \cdot \sum_{j=1}^n (c_{j,k}^q \cdot w_j^q) + \sum_{k=1}^{\dim} \sum_{i=1}^m (c_{i,k}^p \cdot w_i^p) \cdot \sum_{j=1}^m (c_{j,k}^p \cdot w_j^p) \\
 &\quad - 2 \sum_{k=1}^{\dim} \sum_{i=1}^n (c_{i,k}^q \cdot w_i^q) \cdot \sum_{j=1}^m (c_{j,k}^p \cdot w_j^p) \\
 &= \sum_{k=1}^{\dim} (\bar{c}_k^q \cdot \bar{c}_k^q + \bar{c}_k^p \cdot \bar{c}_k^p - 2 \cdot \bar{c}_k^q \cdot \bar{c}_k^p) = \sum_{k=1}^{\dim} (\bar{c}_k^q - \bar{c}_k^p)^2 \\
 &= L_2^2(\bar{c}^q, \bar{c}^p).
 \end{aligned}$$

Consequently, we obtain that L_2 -SQFD(S^q, S^p) is the same as $L_2(\bar{c}^q, \bar{c}^p)$, for any two feature signatures S^q and S^p .

In the proof of Theorem 1, we carry out the following steps: first, we expand the squared Euclidean distance L_2^2 between the corresponding centroids of the feature signatures. Second, we rearrange the terms and make use of the equivalence of the following terms with different indices

$$\begin{aligned}
 -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_i^q w_j^q \cdot \sum_{k=1}^{\dim} c_{i,k}^q \cdot \sum_{k=1}^{\dim} c_{j,k}^q &= -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_i^q w_j^q \cdot \sum_{k=1}^{\dim} c_{j,k}^q \cdot \sum_{k=1}^{\dim} c_{j,k}^q \text{ and} \\
 -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_i^p w_j^p \cdot \sum_{k=1}^{\dim} c_{i,k}^p \cdot \sum_{k=1}^{\dim} c_{j,k}^p &= -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n w_i^p w_j^p \cdot \sum_{k=1}^{\dim} c_{j,k}^p \cdot \sum_{k=1}^{\dim} c_{j,k}^p,
 \end{aligned}$$

and the feature signatures' properties of normalized weights, i.e. $\sum_{i=1}^n w_i^q = \sum_{i=1}^m w_i^p = 1$ which will summarize the indicated lines in the proof to a value of zero. Finally, by further aggregating and rearranging of terms, we finish the proof with the squared Euclidean distance between the weighted mean centroids \bar{c}^q and \bar{c}^p of feature signatures S^q and S^p , respectively.

In Theorem 1, we have shown that the L_2 -Signature Quadratic Form Distance can be computed efficiently between two feature signatures S^q and S^p

by computing the Euclidean distance between their weighted mean centroids \bar{c}^q and \bar{c}^p . These weighted means aggregate the objects' local properties which are exhibited via the corresponding centroids of the feature signatures. Thus, each weighted mean exhibits a single feature signature comprising exactly one centroid with the weight value of one.

By simplifying the computation of the L_2 -Signature Quadratic Form Distance and thus replacing it with an efficient Euclidean distance computation, we are able to improve the efficiency of similarity computations arising in content-based similarity queries inside voluminous multimedia databases. The time complexity and even space complexity of the L_2 -Signature Quadratic Form Distance only depends on the dimensionality of the weighted mean centroids within the underlying feature space. Precomputing and storing the weighted means of the database's feature signatures allows for very efficient similarity computations, but comes at the cost of the similarity measure's expressiveness. While the conventional Signature Quadratic Form Distance can be used with any kind of similarity function and thus exhibits a high expressiveness, the expressiveness of the proposed L_2 -Signature Quadratic Form Distance is limited to that of the Euclidean distance between the corresponding weighted means of the feature signatures.

However, in the next section we will show that the retrieval performance of the proposed L_2 -Signature Quadratic Form Distance is comparable to that of the other state-of-the-art similarity measures: on average, the retrieval quality will not fall below 85%.

4 Experimental Evaluation

In this section, we evaluate the retrieval performance of the L_2 -Signature Quadratic Form Distance in terms of efficiency and effectiveness.

For this purpose, we extracted feature signatures of different size and structure from the following image databases: the *Wang* database [20], the *Coil100* database [13], the *MIR Flickr* database [8], and the *101objects* database [6]. Based on the extracted feature signatures which are generated on a seven-dimensional feature space comprising two position, three color, and two texture dimensions consisting of contrast and coarseness, we randomly queried each database up to 2,000 times with different images varying in their size and structure and measured the *mean average precision* values [11] in order to determine the retrieval performance in terms of effectiveness. More details about the feature signature extraction process and setup can be found in our previous works [23].

We report the results in terms of mean average precision values of the different image databases in Table 1 where we compare the quality of the proposed L_2 -Signature Quadratic Form Distance (L_2 -SQFD) with that of the Hausdorff Distance (*HD*), Perceptually Modified Hausdorff Distance (*PMHD*), Weighted Correlation Distance (*WCD*), Earth Mover's Distance (*EMD*), and Signature Quadratic Form Distance ($SQFD_{f_g}$) based on a Gaussian similarity function [3]. We highlight the highest mean average precision value of each database and expressed the quality of the L_2 -Signature Quadratic Form Distance on a percentage

Table 1. Retrieval performance in terms of mean average precision values of the proposed L₂-Signature Quadratic Form Distance (L₂-SQFD) compared to that of the other state-of-the-art similarity measures

	image database				
	Wang [20]	Coil100 [13]	MIR Flickr [8]	101objects [6]	average
<i>HD</i>	0.308	0.425	0.307	0.072	0.278
<i>PMHD</i>	0.476	0.606	0.322	0.105	0.377
<i>WCD</i>	0.591	0.726	0.335	0.117	0.442
<i>EMD</i>	0.598	0.710	0.333	0.141	0.446
<i>SQFD_{f_q}</i>	0.613	0.776	0.343	0.139	0.468
L ₂ -SQFD	0.566	0.608	0.330	0.094	0.400
quality	92.3%	78.4%	96.2%	66.7%	85.5%

Table 2. Computation time values in milliseconds of the proposed L₂-Signature Quadratic Form Distance (L₂-SQFD) based on the Euclidean distance implementation compared to those of the other state-of-the-art similarity measures

	size of the image database				
	10k	100k	500k	1M	average
<i>HD</i>	51.8	517.5	2534.1	5663.0	2191.6
<i>PMHD</i>	75.0	723.4	3706.3	9204.4	3427.3
<i>WCD</i>	137.8	1332.9	6810.9	14690.9	5743.1
<i>EMD</i>	473.7	4666.3	23140.4	50445.1	19681.4
<i>SQFD_{f_q}</i>	299.4	2884.0	14579.9	29820.4	11895.9
L ₂ -SQFD	4.4	49.0	331.5	840.3	306.3
avg. speed-up factor	47.2	41.3	30.6	26.1	28.0

basis of these values. It can be seen in the table, that the best mean average precision values are always higher than the mean average precision values of the L₂-Signature Quadratic Form Distance. However, on average our proposed approach reaches a high retrieval performance in terms of effectiveness of greater than 85%. The best performance of 96.2% is obtained in the MIR Flickr database while the lowest performance of 66.7% is obtained in the 101objects database. Moreover, the L₂-Signature Quadratic Form Distance exhibits even higher mean average precision values than those of the Hausdorff Distance and the Perceptually Modified Hausdorff Distance.

Regarding effectiveness, we compared the L₂-Signature Quadratic Form Distance implemented as Euclidean distance, as stated in Theorem 1 with the other similarity measures on a large-scale image database containing up to one million images. We measured the computation time values needed to generate complete rankings of the image databases on a 2.4 GHz Intel Core 2 Duo machine with 2 GB main memory and implemented all approaches in JAVA 1.6.

In Table 2, the computation time values are given in milliseconds. It can be seen in the table that the computation time values of the proposed L₂-Signature Quadratic Form Distance implemented as Euclidean distance are significantly lower than those of the other similarity measures. The L₂-Signature

Quadratic Form Distance completes the ranking of the $10k$ image database in approximately four milliseconds and that of the $1M$ image database in less than one second. As a result, our approach reaches the minimum average speed-up factor of 26 in the $1M$ image database and the maximum average speed-up factor of 47 in the $10k$ image database. On average, the L_2 -Signature Quadratic Form Distance is able to finish a ranking of the complete database 28 times faster than the other similarity measures can do.

To sum up, we have shown that the L_2 -Signature Quadratic Form Distance can be computed very efficiently by the Euclidean distance among the feature signatures' weighted mean centroids. As a result, our approach reduces the computation time spent for costly content-based similarity computations in voluminous multimedia databases by a factor up to 47, while maintaining on average a high retrieval quality of greater than 85% compared to the other state-of-the-art similarity measures. Thus, we conclude that our approach is able to process millions of similarity computations in less than a few seconds while maintaining a high retrieval quality.

5 Conclusions and Outlook

In this paper, we introduced the L_2 -Signature Quadratic Form Distance as a specific instance of the Signature Quadratic Form Distance. By showing its efficient Euclidean-based implementation, we tackled the efficiency problem of content-based searching for similar objects with adaptive similarity measures inside voluminous multimedia databases. The proposed approach improves the computation time spent for the immense amount of costly similarity computations among the stored data objects by a factor up to 47 while maintaining a high retrieval quality of greater than 85% on average. As a result, we are able to process millions of similarity computations in less than a few seconds.

As for future work, we plan to examine the applicability of the other state-of-the-art similarity measures to very large multimedia databases, as well as their expressiveness. Furthermore, we plan to investigate more compact content representation forms of multimedia data.

References

1. Beecks, C., Uysal, M.S., Seidl, T.: Signature Quadratic Form Distances for Content-Based Similarity. In: Proc. of ACM Int. Conf. on Multimedia, pp. 697–700 (2009)
2. Beecks, C., Uysal, M.S., Seidl, T.: A comparative study of similarity measures for content-based multimedia retrieval. In: Proc. of IEEE Int. Conf. on Multimedia and Expo., pp. 1552–1557 (2010)
3. Beecks, C., Uysal, M.S., Seidl, T.: Signature Quadratic Form Distance. In: Proc. of ACM Int. Conf. on Image and Video Retrieval, pp. 438–445 (2010)
4. Deselaers, T., Keysers, D., Ney, H.: Features for Image Retrieval: An Experimental Comparison. Information Retrieval 11(2), 77–107 (2008)

5. Faloutsos, C., Barber, R., Flickner, M., Hafner, J., Niblack, W., Petkovic, D., Equitz, W.: Efficient and Effective Querying by Image Content. *Journal of Intelligent Information Systems* 3(3/4), 231–262 (1994)
6. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples an incremental bayesian approach tested on 101 object categories. In: Proc. of the Workshop on Generative-Model Based Vision (2004)
7. Hu, R., Rüger, S.M., Song, D., Liu, H., Huang, Z.: Dissimilarity measures for content-based image retrieval. In: Proc. of IEEE Int. Conf. on Multimedia and Expo., pp. 1365–1368 (2008)
8. Huiskes, M.J., Lew, M.S.: The MIR Flickr Retrieval Evaluation. In: Proc. of ACM Int. Conf. on Multimedia information retrieval, pp. 39–43 (2008)
9. Huttenlocher, D., Klanderman, G., Rucklidge, W.: Comparing images using the Hausdorff Distance. *IEEE Trans. Pattern Anal. Mach. Intell.* 15(9), 850–863 (1993)
10. Leow, W.K., Li, R.: The analysis and applications of adaptive-binning color histograms. *Computer Vision and Image Understanding* 94(1-3), 67–91 (2004)
11. Manning, C.D., Raghavan, P., Schütze, H.: *Introduction to Information Retrieval*. Cambridge University Press, Cambridge (2008)
12. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.* 27(10), 1615–1630 (2005)
13. Nene, S., Nayar, S.K., Murase, H.: Columbia Object Image Library (COIL-100). Tech. rep., Department of Computer Science, Columbia University (1996)
14. Park, B.G., Lee, K.M., Lee, S.U.: Color-based image retrieval using perceptually modified Hausdorff distance. *Journal on Image and Video Processing*, 1–10 (2008)
15. Rubner, Y.: Perceptual metrics for image database navigation. Ph.D. thesis (1999)
16. Rubner, Y., Tomasi, C., Guibas, L.J.: The Earth Mover's Distance as a Metric for Image Retrieval. *Int. Journal of Computer Vision* 40(2), 99–121 (2000)
17. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 32(9) (2010)
18. Seidl, T., Kriegel, H.-P.: Efficient User-Adaptable Similarity Search in Large Multimedia Databases. In: Proc. of Int. Conf. on Very Large Data Bases, pp. 506–515 (1997)
19. Veltkamp, R., Tanase, M., Sent, D.: Features in content-based image retrieval systems: A survey. *State-of-the-art in content-based image and video retrieval*, 97–124 (2001)
20. Wang, J.Z., Jia, L., Wiederhold, G.: SIMPLIcity: semantics-sensitive integrated matching for picture libraries. *IEEE Trans. Pattern Anal. Mach. Intell.* 23(9), 947–963 (2001)

Generating Representative Views of Landmarks via Scenic Theme Detection

Yi-Liang Zhao¹, Yan-Tao Zheng², Xiangdong Zhou³, and Tat-Seng Chua¹

¹ Department of Computer Science, National University of Singapore, Singapore

² Institute for Infocomm Research, Singapore

³ Fudan University, China

{zhaoyl,chuats}@comp.nus.edu.sg, yantaozheng@gmail.com,
xdzhou@fudan.edu.cn

Abstract. Visual summarization of landmarks is an interesting and non-trivial task with the availability of gigantic community-contributed resources. In this work, we investigate ways to generate representative and distinctive views of landmarks by automatically discovering the underlying *Scenic Themes* (e.g. sunny, night view, snow, foggy views, etc.) via a content-based analysis. The challenge is that the task suffers from the subjectivity of the scenic theme understanding, and there is lack of prior knowledge of scenic themes understanding. In addition, the visual variations of scenic themes are results of joint effects of factors including weather, time, season, etc. To tackle the aforementioned issues, we exploit the Dirichlet Process Gaussian Mixture Model (DPGMM). The major advantages in using DPGMM is that it is fully unsupervised and do not require the number of components to be fixed beforehand, which avoids the difficulty in adjusting model complexity to avoid over-fitting. This work makes the first attempt towards generation of representative views of landmarks via scenic theme mining. Testing on seven famous world landmarks show promising results.

Keywords: Dirichlet Process, Dirichlet Process Gaussian Mixture Model, Scenic Theme Detection.

1 Introduction

The fast development and proliferation of digital photo-capture devices and the growing practice of online photo-sharing have resulted in huge online photo collections, which cover virtually everywhere on earth. The fast-growing of this huge image collection opens up many opportunities to research communities to work on more effective and efficient searching, viewing, archiving and interaction with such collections. Recently, much work aims to organize this huge collection or mine important landmarks worldwide based on the context information: geography, user, tag, etc. [5] [9] [19]. However, less efforts have been put into generating representative views of landmarks via recognizing the underlying scenic themes. As shown in Figure 1, a landmark may present different scenery views under different time, season and weather circumstances. Here, we define a distinct landmark scenery view as a scenic theme of the landmark. A scenic theme affords distinct vistas of the landmark with notably different aesthetic and visual

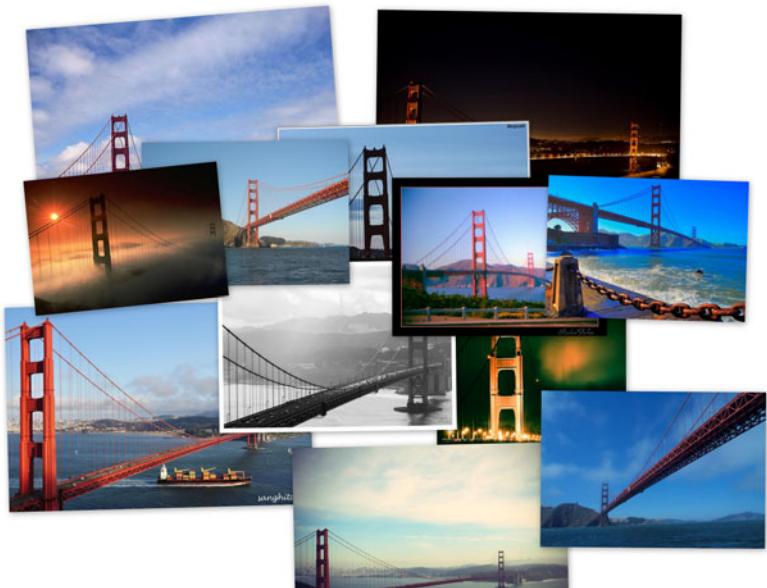


Fig. 1. Different scenic themes of Golden Gate Bridge

perceptions. Effective scenic theme detection helps to better organize, browse and index image collections of a particular landmark. In particular, we show in this work that the summary of distinctive and representative views is generated with satisfaction on top of the detected scenic themes.

However, mining representative scenic themes from community-contributed image collections itself is a difficult task. Variations of scenic themes are the results of joint effects of seasons, weather conditions and time. Although we have the corresponding context information associated with the images most of the time, it is very difficult to model the underlying scenic themes using a combination of these meta information. In addition, scenic theme understanding requires subjective judgements. People from different background, culture and with different personal experience may have different perceptions on the same scenes. Figure 2 shows the variations of scenic views of The Eiffel Tower in a cloudy or overcast day. People may judge that it is either cloudy or overcast based on their own experience and background. In addition, we find that landmarks in different parts of the world often have different sets of distinctive scenic themes. For example, in tropical area, there is probably no snowy scenes but lots of beautiful night views whereas in subpolar zones, we can find nice snowscapes. After all, we do not have a fixed list of scenic themes, which makes automatic detection and generation of scenic themes summary a necessity. To tackle this problem, we adopt a generative probabilistic approach to visually model the scenic themes of a landmark. The rational is that scenic theme of a landmark is a cause-effect process and the generative model is expected to capture the underlying rules of producing different landmark sceneries.



Fig. 2. Scene-variation views of The Eiffel Tower in a cloudy or overcast day

The task of mining distinct scenic themes is formulated as a clustering problem in this paper. Traditional probabilistic models are often used throughout machine learning to model the distributions over observed data, but they tend to suffer from either under-fitting or over-fitting. The choice of using bayesian nonparametric approach resolves both under-fitting by using a model with an unbounded complexity and over-fitting by approximating the full posterior over parameters. Dirichlet Process (DP) is currently one of the most popular non-parametric bayesian model [7] [2] [6] [12] [14] due to its simplicity and efficiency. Here we show that the Gaussian mixtures whose parameters are generated by DP is able to create satisfactory view summaries with distinctive scenic themes.

In summary, the contribution of this paper is: we demonstrate a novel attempt in generating view summaries via underlying scenic themes detection on community-contributed images by using DPGMM. Testing on the seven landmarks shows that the proposed approach delivers promising results. The remaining parts of the paper is organized as follows. In Section 2, we review the related work. Formulation of the problem together with the proposed model are presented in Section 3. We present the experiments in Section 4 followed by the conclusion in Section 5.

2 Related Work

We review two areas, which are most related to our work: (1) Visual scene understanding; (2) Visual summarization through Landmark mining from community-contributed collections.

The use of global features for scene classification dates back in 1998, when Martin, et al. [18] showed how global features can be used to model each scene as an individual object for classification. Their approaches, however, are normally only used to classify a small number of scene categories. Recently, probabilistic models are used extensively in visual scenes categorization in the computer vision literature [10][17][4][8]. Fei-Fei, et al. [10] proposed a bayesian hierarchical model for learning natural scene categories based on Latent Dirichlet Allocation (LDA). However, a fixed list of categories are readily available in their work while we need to tackle the problem of variable number of scenic themes for different landmarks.

Visual summarization through landmark mining from online community-contributions is a recent trend [3][9][1][19][16]. Lyndon, et al. [9] proposed a way to generate representative views of important landmarks based on both context and content information. The statistical approaches adopted by them showed effectiveness in aggregating the representative views of landmarks in San Francisco area. However, statistical approach needs a highly accurate and sufficiently large dataset. Simon, et al. [16] worked on finding a set of canonical views to summarize a visual scene. Their work aimed at constructing a guidebook which contains a summary on representative views of large landmarks. Comparing to their work, ours focuses on generating a representative view summarization distinguished by the underlying scenic themes, which is more difficult due to its subjectivity and uncertainty. Zheng, et al. [19] built a landmark recognition engine in modeling and recognizing landmarks at world-scale level. The graph clustering result however only shows strong visual correlations between the traditional representative views of landmarks. While these work did not look into ways to organize the images based on scenic themes, the approaches adopted in mining representative landmarks, however, can be used in the preprocessing steps of our work in generating clearer subsets of the collections.

3 Our Approach

In this section, we formally define the problem and elaborate on details of the proposed approach.

3.1 Problem Formulation

Let \mathbf{y} denote the scenery or visual appearance of a landmark. \mathbf{y} is the consequence of joint effects of factors \mathbf{q} , like weather, season, lighting, etc, with markov condition $\mathbf{q} \rightarrow \mathbf{y}$. To model this cause-effect process is, however, a challenge, as the causality relationship is nondeterministic and not tractable with a finite number of rules and parameters. To simplify the modeling, we introduce an intermediate variable \mathbf{t} , i.e., scenic theme. In the new cause-effect process, scenic theme is a result of joint effects of factors like weather, season, etc, while landmark scenery (visual appearance) is an observation conditioned on scenic theme only. The markov condition now becomes $\mathbf{q} \rightarrow \mathbf{t} \rightarrow \mathbf{y}$. Scenic theme

here corresponds to a distribution of random variable \mathbf{y} of landmark visual appearances. Intuitively, a scenic theme captures the characteristics of landmark scenery appearances from certain aspect. For example, day view and night view could be examples of scenic theme, while day view can be subdivided further. As our goal is to generate visual summarization of landmarks, we focus only on the second half of the cause-effect process, i.e., $t \rightarrow \mathbf{y}$ and model it using a generative probabilistic model. To avoid suffering from the problem of over-fitting or under-fitting associated with traditional parametric models when there is a mismatch between the complexity of the model and the amount of the data available, we propose to use the bayesian nonparametric approach for the clustering problem. According to Rasmussen et al. [15], reasonable and proper bayesian methods do not have over-fitting problem as the number of latent variables does not grow with the number of mixture components in the model inference. Formally, given observations: $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ with $\mathbf{y}_i \sim G$ drawn from some unknown distribution G , we first place a prior over G then compute the posterior over G given the observations. In this case, we use Dirichlet Process Gaussian Mixture Model (DPGMM) as the prior distribution of the model since DP has a tractable posterior distribution [7][14]. The nonparametric nature of the DP translates to mixture models with a countably infinite number of components. Use of countably infinite gaussian mixtures bypasses the need to determine the "correct" number of components in a finite mixture model, which is technically much more difficult.

In generating view summaries based on the detected scenic themes, we seek to find the most distinct and representative scenic theme clusters. The detected scenic themes are each modeled as a probability distribution; in this case, each gaussian component represents a distinct detected scenic themes. We then seek to find the differences between each detected scenic themes by calculating the Kullback-Leibler (KL) Divergence between each corresponding probability distributions. The KL Divergence between two Gaussian distributions: $\mathcal{N}_i(\boldsymbol{\mu}_i, \Sigma_i)$ and $\mathcal{N}_j(\boldsymbol{\mu}_j, \Sigma_j)$ is:

$$D_{KL}(\mathcal{N}_i || \mathcal{N}_j) = \frac{1}{2} \left(\ln \left(\frac{|\Sigma_j|}{|\Sigma_i|} \right) + \text{tr}(\Sigma_j^{-1} \Sigma_i) + (\boldsymbol{\mu}_j - \boldsymbol{\mu}_i)^T \Sigma_j^{-1} (\boldsymbol{\mu}_j - \boldsymbol{\mu}_i) - N \right) \quad (1)$$

For each landmark, we rank the detected scenic themes according to the sum of KL Divergence values with respect to other theme clusters and select the top ten themes to form the visual summary of the landmark. The overall framework is presented in Figure 3.

3.2 Dirichlet Process Gaussian Mixture Model (DPGMM)

As one of the bayesian non-parametric model, DPGMM does not require the number of Gaussian components to be fixed in advance. Instead, the number of components is determined by the model and data in the subsequent inferences. The infinite DPGMM for scenic themes detection is defined as follows: we model a set of observations: $\{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ using a set of latent parameters $\{(\boldsymbol{\mu}_1, \mathbf{S}_1), \dots, (\boldsymbol{\mu}_n, \mathbf{S}_n)\}$, where $\boldsymbol{\mu}_i$ are the means, \mathbf{S}_i are the precisions (inverse

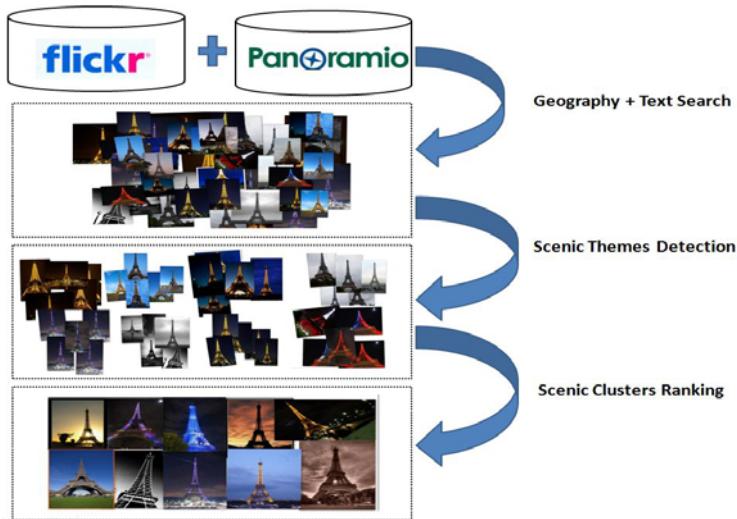


Fig. 3. Overall Framework

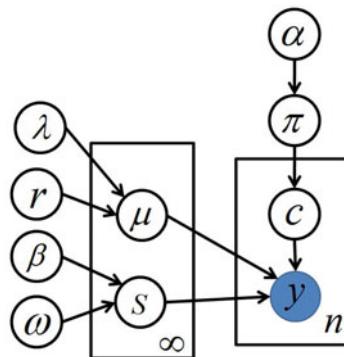


Fig. 4. Dirichlet Process Gaussian Mixture Model: y is the observation and c is the class label

variances), both of which are generated by a DP which can be seen as a distribution over the parameters of other distributions. In this case, each μ_i is drawn independently and identically from another Gaussian distribution parameterized by the hyperparameters: λ and r and each S_i is drawn independently and identically from a Gamma distribution parameterized by β and ω . Together, S and μ form the scenic theme intermediate variable t as introduced in Section 3.1. Let c_i be a cluster assignment variable, which takes on value k with probability π_k . Thus each y_i has distribution $p(y|\mu, S, c)$ parameterized by μ, S and c . The graphical representation of this model is presented in Figure 4 and is defined as follows:

$$\begin{aligned}
\boldsymbol{\pi}|\alpha &\sim \text{Dirichlet}\left(\frac{\alpha}{k}, \dots, \frac{\alpha}{k}\right) = \frac{\Gamma(\alpha)}{\Gamma(\frac{\alpha}{k})^k} \prod_{j=1}^k \pi_j^{\frac{\alpha}{k}-1} \\
\mathbf{c}|\boldsymbol{\pi} &\sim \text{Multinomial}(\boldsymbol{\pi}) = \prod_{j=1}^k \pi_j^{n_j} \\
\boldsymbol{\mu}|\boldsymbol{\lambda}, \mathbf{r} &\sim \mathcal{N}(\boldsymbol{\lambda}, \mathbf{r}^{-1}) \\
\mathbf{S}|\boldsymbol{\beta}, \boldsymbol{\omega} &\sim \text{Gamma}(\boldsymbol{\beta}, \boldsymbol{\omega}^{-1}) \\
p(\mathbf{y}|\boldsymbol{\mu}, \mathbf{S}, \mathbf{c}) &= \sum_{j=1}^k \pi_j \mathcal{N}(\boldsymbol{\mu}_j, \mathbf{S}_j^{-1})
\end{aligned} \tag{2}$$

In equation (2), n_j is the number of images belonging to the j th scenery cluster. The indicator variables \mathbf{c} is introduced to encode which class has generated the observation so that the inference is possible using finite amounts of computation with the maximum number of components not exceeding the number of observations. Observations: \mathbf{y}_i are color histogram feature vectors in the current setting. The mixing proportions $\boldsymbol{\pi}$ are positive and sum to one. From equation (2), we can write the prior directly in terms of the indicators by integrating out the mixing proportions:

$$\begin{aligned}
p(c_1, \dots, c_k|\alpha) &= \int p(c_1, \dots, c_k|\pi_1, \dots, \pi_k) d\pi_1 \cdots d\pi_k \\
&= \frac{\Gamma(\alpha)}{\Gamma(n+\alpha)} \prod_{j=1}^k \frac{\Gamma(n_j + \alpha/k)}{\Gamma(\alpha/k)}
\end{aligned} \tag{3}$$

With the analytical tractability of equation (3), we can now work directly with the finite number of indicator variables, rather than the infinite number of mixing proportions. We use Markov Chain which relies on Gibbs Sampling for inference in this work [13]. Each variable is updated by sampling from its posterior distribution conditional on all other variables. We repeatedly sample the parameters, hyperparameters and indicator variables from their posterior distributions conditioned on all other variables: (1) sample parameters conditioned on the indicators and hyperparameters; (2) sample hyperparameters conditioned on the parameters; (3) update each indicator, conditioned on other indicators and hyperparameters; and (4) update Dirichlet process concentration parameter α . The process will repeat until termination condition is met. In this paper, we set the maximum number of iteration to be 5,000 for each landmark.

4 Experiment

The experiments are performed on seven worldwide famous landmarks: The Eiffel Tower, Golden Gate Bridge, The Great Sphinx, Notre Dame, Leaning Tower Pisa, Statue of Liberty and Basil Cathedral. To make the dataset more complete, we crawl data from two online collections: Flickr¹ and Panoramio². Comparing with Panoramio, which is a geolocation-oriented photo sharing website, Flickr has much more user contribution while Panoramio images have more accurate geographical locations. Using public APIs provided by both web services, we

¹ <http://www.flickr.com>

² <http://www.panoramio.com/>

specify a restricted bounding box on geographical locations as well as key words related to corresponding landmarks. After cleaning the data, we have an average of 806 images of each landmark. In the scenic theme detection stage, we exploit global features in the work due to its capability in producing compact representations of images. We did a comparison test with different global features related to the color distributions: Color Histogram, Color Moment and Color Correlogram. We found that color histogram gives the best results in generating the most coherent scenic theme classes. After feature extraction, we normalize the feature vectors such that the sum of each element equals to a constant: $\sum_k y_{ik} = \sum_k y_{jk} = a$, for every $i \neq j$. We then adopt Principle Component Analysis (PCA) approach to reduce the dimensionality of the feature vectors such that: $d_{new} = [b \times \sqrt{n}]$, where d_{new} is the reduced dimension and n is the number of images in the dataset for that particular landmark. Empirically, we choose $b = 2.5$ in the current setting because it produces better results compared to other values. After sampling from the posterior distribution of the DPGMM according to the Markov Chain inference procedure described in Section 3, we obtained an average of 15 scenic themes for each landmarks. Some detected scenic themes of The Eiffel Tower are presented in Figure 5. Finally, to give a view summarization with the most interesting and representative scenic themes for each landmark, we exploit the calculated probability measures of each mixture model and measure the pair-wise KL Divergences according to equation (1). We rank the detected scenic themes according to the sum of KL Divergence values with respect to other theme clusters and select the top ten themes to form the visual summary of each landmark. We then calculate the probability

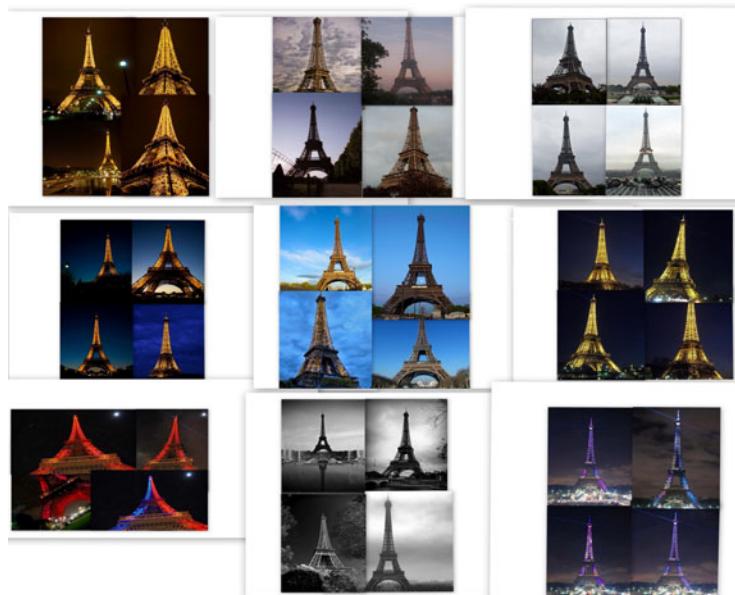


Fig. 5. Scenic themes mined for The Eiffel Tower

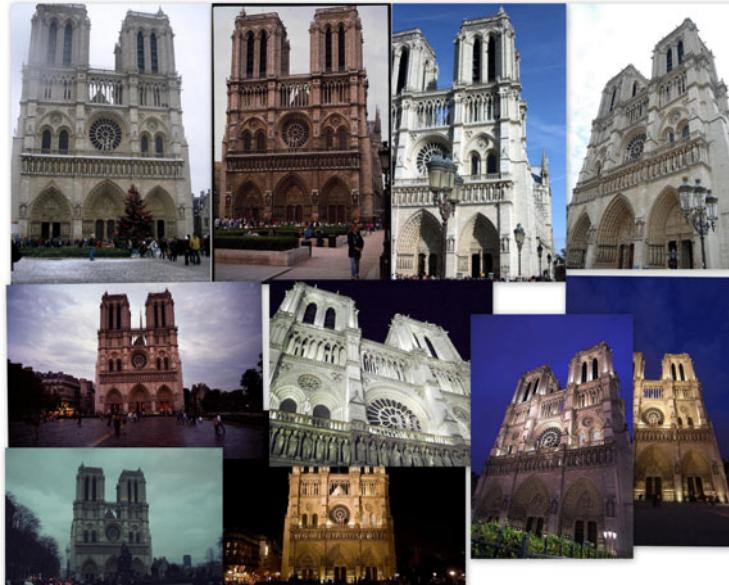


Fig. 6. View summarization with representative scenes detected for Notre Dame

of images in each selected theme clusters and choose the one with the highest probability to represent the theme. Figure 6 shows the view summary of Notre Dame generated by our approach.

To evaluate the correctness and representativeness of the mined scenic themes, we have conducted a user study, where a group of users are selected to judge the mined scenic themes for each landmarks. The judgments are based on four criteria: (1) the level of consistency of the scenic themes mined for each landmark; (2) the level of completeness of the generated scenic themes summary; (3) the level of redundancy of the summary and (4) how satisfying are the summaries? We randomly selected twenty users: seven from Singapore; others from Shanghai, China. The evaluation result is depicted in Figure 7. The results for each question are averaged over all users for each landmark. The average satisfaction score is 7.97. We observe that Statue of Liberty and Basil Cathedral obtained the best scores while The Great Sphinx does not perform well in terms of the consistency level. The reasons could be: (1) there are much fewer distinct scenic themes The Great Sphinx has as compared to that of The Statue of Liberty and Basil Cathedral; and (2) the color distribution of The Great Sphinx is very similar to most of the background (i.e. the desert). To mitigate this problem, we could look into extracting regions mostly related to the scenic themes by using camera calibration technique and incorporating spatial information. In addition, Basil Cathedral scored lowest in both uniqueness and completeness measures. We attribute the low scores to the mismatch between the availability of sufficient data and the expectations of more nice views from the users. In summary, our

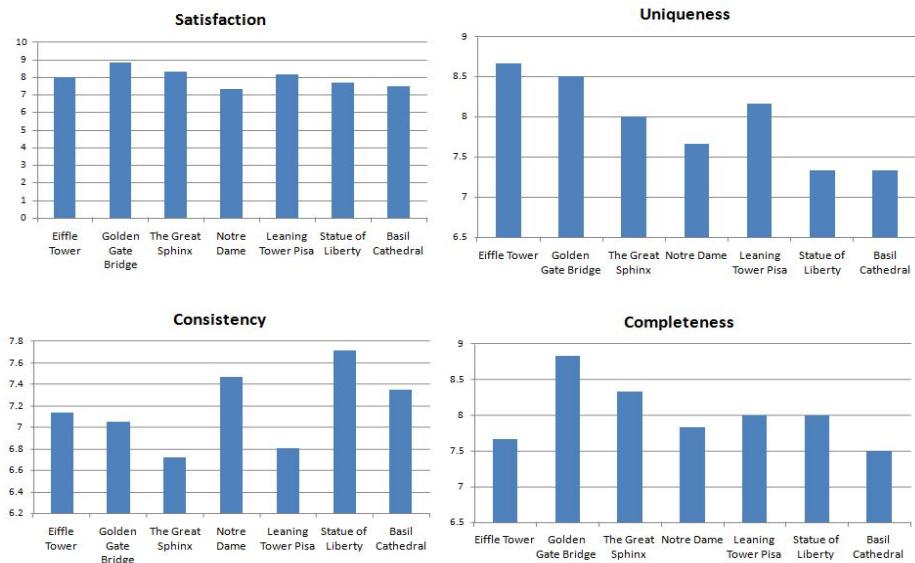


Fig. 7. Average scores for each of the four evaluation questions

proposed model yields satisfactory results and more efforts are required to tackle the problems caused by data sparsity and noises contributed by various factors.

5 Conclusion

We have presented a novel attempt to use Dirichlet Process Gaussian Mixture Model (DPGMM) to generate visual summarization of landmarks via scenic theme detection. Our approach shows promising results in generating satisfactory scenic themes for most of the landmarks. A user study is done with good response in terms of the representativeness and coherence of the scenic theme clusters. In the future, we shall look into building applications which allow users to browse image collections organized by distinct scenery views. In addition, contextual information could be exploited to boost the performance when more accurate meta information and richer web services are available [11].

References

1. Ahern, S., Naaman, M., Nair, R., Yang, J.H.-I.: World explorer: visualizing aggregate data from unstructured text in geo-referenced collections. In: JCDL, pp. 1–10 (2007)
2. Antoniak, C.E.: Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems. *The annals of statistics* 2(6), 1152–1174 (1974)
3. Berg, T.L., Forsyth, D.A.: Automatic ranking of iconic images. Technical report (2007)

4. Zisserman, A., Bosch, A., Munoz, X.: Scene classification via pLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) *ECCV 2006*. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)
5. Crandall, D.J., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: *WWW*, pp. 761–770 (2009)
6. Escobar, M.D., West, M.: Bayesian Density Estimation and Inference Using Mixtures.. *Journal of the american statistical association* 90(430) (1995)
7. Ferguson, T.S.: A Bayesian analysis of some nonparametric problems. *The annals of statistics* 1(2), 209–230 (1973)
8. Fritz, M., Schiele, B.: Decomposition, discovery and detection of visual categories using topic models. In: *CVPR*, vol. 0, pp. 1–8 (2008)
9. Kennedy, L.S., Naaman, M.: Generating diverse and representative image search results for landmarks. In: *WWW*, pp. 297–306 (2008)
10. Li, F.-F., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: *CVPR*, vol. 2, pp. 524–531 (2005)
11. Naaman, M., Harada, S., Wang, Q., Garcia-Molina, H., Paepcke, A.: Context data in geo-referenced digital photo collections. In: *MM*, pp. 196–203 (2004)
12. Neal, R.M.: Bayesian mixture modeling. In: *Maximum Entropy and Bayesian Methods: Proceedings of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, Seattle, pp. 197–211 (1991)
13. Neal, R.M.: Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics* 9(2), 249–265 (2000)
14. Rasmussen, C.E.: The infinite Gaussian mixture model. *Advances in neural information processing systems* 12, 554–560 (2000)
15. Rasmussen, C.E., Ghahramani, Z.: Occam razor. In: *Advances in neural information processing systems 13: Proceedings of the 2000 Conference*, p. 294. The MIT Press, Cambridge (2001)
16. Simon, I., Snavely, N., Seitz, S.M.: Scene summarization for online image collections. In: *ICCV*, pp. 1–8. Citeseer (2007)
17. Sudderth, E.B., Torralba, A., Freeman, W.T., Willsky, A.S.: Learning hierarchical models of scenes, objects, and parts. In: *ICCV*, vol. 2, pp. 1331–1338 (2005)
18. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: *Proceedings of 1998 IEEE International Workshop on Content-Based Access of Image and Video Database*, pp. 42–51 (1998)
19. Zheng, Y.-T., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T.-S., Neven, H.: Tour the world: building a web-scale landmark recognition engine. In: *CVPR* (June 2009)

Regularized Semi-supervised Latent Dirichlet Allocation for Visual Concept Learning

Liansheng Zhuang^{1,2}, Lanbo She², Jingjing Huang², Jiebo Luo³, and Nenghai Yu^{1,2}

¹ MOE-MS Keynote Lab of MCC, USTC, Hefei 230027, China

² School of Information Science and Technology, USTC, Hefei 230027, China

³ Kodak Research Labs, Eastman Kodak Company, Rochester, New York 14650, USA

{lszhuang, ynh}@ustc.edu.cn, jiebo.luo@kodak.com

Abstract. Topic models are a popular tool for visual concept learning. Current topic models are either unsupervised or fully supervised. Although lots of labeled images can significantly improve the performance of topic models, they are very costly to acquire. Meanwhile, billions of unlabeled images are freely available on the internet. In this paper, to take advantage of both limited labeled training images and rich unlabeled images, we propose a novel technique called regularized Semi-supervised Latent Dirichlet Allocation (r-SSLDA) for learning visual concept classifiers. Instead of introducing a new topic model, we attempt to find an efficient way to learn topic models in a semi-supervised way. r-SSLDA considers both semi-supervised properties and supervised topic model simultaneously in a regularization framework. Experiments on Caltech 101 and Caltech 256 have shown that r-SSLDA outperforms unsupervised LDA, and achieves competitive performance against fully supervised LDA, while sharply reducing the number of labeled images required.

Keywords: Visual Concept Learning, Latent Dirichlet Allocation, Semi-supervised Learning.

1 Introduction

Visual concept detection is a basic problem in many applications such as image retrieval. It aims at automatically mapping images into predefined semantic concepts (such as indoor, sunset, airplane, face, etc.), so as to bridge the so-called semantic gap between low-level visual features and high-level semantic content of images. Although it has been studied for many years, it is still a challenging problem within multimedia and computer vision. Learning visual concept classifiers is the key problem to visual concept detection. Recently, topic models such as Latent Dirichlet Allocation (LDA) [6] are popular for solving the problem [1-5]. Topic models, which clusters co-occurring words into topics, are a very efficient and effective tool from the field of text analysis. In recent years, Sivic et al. introduced it into the field of computer vision and multimedia [1-5, 7, 10, 12]. In these applications, images are treated as documents, and represented by a histogram of visual words. A visual word is equivalent to a text word, and it is often generated by clustering various local descriptors such as SIFT [8].

Classic LDA is an unsupervised model without using any prior label information. The lack of useful supervised information usually leads to slow convergence and unsatisfactory performance. Moreover, only the visual words in the training images are modeled in classic LDA. During classification, class labels are simply treated as features extracted from the topic distribution [4, 5]. Since the class label is not part of the model, classic LDA is not well suited for classification problems, thus resulting in not so robust performance in visual concept detection.

To make LDA more effective for classification and prediction problem, Blei et al. introduced the supervised Latent Dirichlet Allocation (sLDA) model [9, 10], in which the label parameter is domain structure and topics are trained to best fit the corresponding variables or labels. Visual words and class labels in the training images are modeled at the same time in sLDA. Experiments showed that sLDA outperforms classic LDA significantly for image classification problems. Similarly, Wang et al. [17] proposed a Semi-Latent Dirichlet Allocation for human action recognition. Different from sLDA, Semi-LDA introduces supervised information into its model by associating image class labels with visual words. That is, Semi-LDA assumes that the topic of a visual word is observable and equal to the image class label. In this way, Semi-LDA achieves better performance than classic LDA. Figure 1 shows the graphic model representation of LDA, sLDA and Semi-LDA.

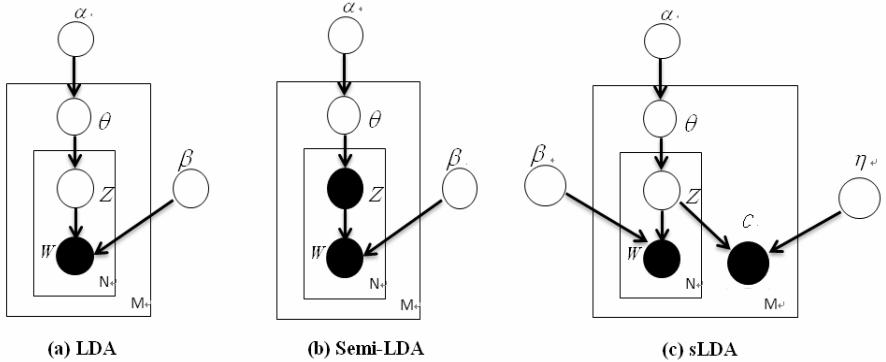


Fig. 1. Graph model representation of classic LDA (a), Semi-LDA (b) and full supervised LDA (c)

However, both sLDA and Semi-LDA improve the performance in a fully supervised fashion, and therefore require all training images to be labeled. For a large dataset, any label information is labor intensive, making sLDA or Semi-LDA greatly restricted to only a few object classes. On the other hand, huge amounts of unlabeled images are available and easy to obtain. These unlabeled images contain enough information to train visual classifiers. Moreover, unlabeled images can help avoid overfitting. Therefore, learning classifiers with a fully supervised topic model in a semi-supervised manner, which aims to utilize a large amount of unlabeled data, is a promising direction to explore.

Although much work on semi-supervised learning (SSL) algorithms has been developed to solve practical problems [11], few considered combining semi-supervised properties with topic models to solve the visual concept detection problem. Recently,

Zhuang etc. [12] proposed a method called Semi-supervised pLSA (Ss-pLSA) for image classification. By introducing category label information into the EM algorithm during training, they can train classifiers with pLSA in a semi-supervised fashion. Experimental results showed that supervised information could effectively speed up the convergence to achieve desire results. But Ss-pLSA seems to be a loosely coupled way of simple label propagation in conjunction with a unsupervised PLSA model. Moreover, it does not encode class labels into its model, similar to classic PLSA.

In this paper, we propose a novel algorithm called regularized Semi-supervised Latent Dirichlet Allocation (r-SSLDA) for visual concept learning. Inspired by [13], instead of attempting to introduce a new Bayesian statistical model, we tried to find an efficient semi-supervised way to learn visual concept classifier with topic models. Unlike the loosely coupled solution in [12], we consider both semi-supervised properties and topic models simultaneously in a regularization framework. By minimizing the cost function of the regularization framework, we provide a direct solution to the semi-supervised topic model problem. Different from Ss-pLSA, our r-SSLDA encodes class labels into its framework by adopting a supervised LDA model to learn the visual concept classifiers. Experimental results showed that r-SSLDA significantly outperformed classic unsupervised LDA and achieved competitive performance compared with sLDA, while drastically reducing the number of labeled images needed. Furthermore, experimental results showed that our r-SSLDA was more effective than simple semi-supervised LDA (s-SSLDA), which simply implements the supervised LDA twice.

The rest of this paper is organized as follows: In Section 2, we will give the detail of the regularized Semi-supervised LDA. Experiments and result analysis follow in Session 3. Section 4 contains the conclusions.

2 Regularized Semi-supervised LDA

In this section, we first present the regularization framework of our semi-supervised topic model problem, and then describe our regularized semi-supervised LDA algorithm in detail.

2.1 Regularization Framework

Give a training image set $X = \{x_1, x_2, \dots, x_n\} \subset \mathcal{R}^d$, where x_i is the $i - th$ image, and n is the total number of training images. Let $y = (y_1, y_2, \dots, y_n)^T$ be the initial label vector, and $y_i \in \{-1, 1\}$ is the label of image x_i . We set y_i to 1 (positive image) or -1 (negative image). For unlabeled images, y_i can be any other limited value. Let F denote the set of $n \times 1$ vectors. A vector $f \in F$ corresponds to a classification function defined on X . $\forall f \in F$ assigns a real value f_i to each image x_i , where f_i is the $i - th$ element of f . The label of the unlabeled image x_u is determined by the sign of f_u . To find the optimal vector f^* to classify X , we design a cost function $Q(f)$ as follows:

$$f^* = \arg \min_f Q(f) = \arg \min_f (Q_{smoothness} + \mu Q_{fitting}^L) \quad (1)$$

The first term $Q_{smoothness}$ is the smoothness cost. It means that a good classification function should not change too much between nearby sample points. That is, images that are close nearby in the feature space (thus similar) tend to have the same labels. Similar to the standard SSL algorithm [14], we define the smoothness cost function as follows:

$$Q_{smoothness} = \frac{1}{2} \sum_{i,j=1}^n W_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2 \quad (2)$$

where W represents the similarity between two images x_i and x_j . d_i is the sum of the i -th row of W .

The second term $Q_{fitting}^L$ means that a good classification function should not change too much from the initial label assignment. So we define the fitting cost as follows:

$$Q_{fitting}^L = \sum_{x_i \in X^L} (f_i - y_i)^2 \quad (3)$$

where X^L means a set of labeled images. Note that, $Q_{fitting}^L$ is only used on the labeled images. For unlabeled images, y_j is indefinite. The regularization parameter μ controls the trade-off between constraints, and is empirically set to 1/9 in our experiments.

Thus, the cost function in our semi-supervised topic model is defined as:

$$Q(f) = \frac{1}{2} \sum_{i,j=1}^n W_{ij} \left(\frac{f_i}{\sqrt{d_i}} - \frac{f_j}{\sqrt{d_j}} \right)^2 + \mu \sum_{x_i \in X^L} (f_i - y_i)^2 \quad (4)$$

2.2 Regularized Semi-supervised LDA Algorithm

To minimize equation (4) with respect to f , we assume that the affinity matrix W is symmetric and irreducible (?). Let D be a diagonal matrix with its (i,i) -element equals to the sum of the i -th row of W . Therefore, we rewrite the cost function as:

$$Q(f) = f^T (I - D^{-1/2} W D^{1/2}) f + \mu (f - y)^T I^L (f - y) \quad (5)$$

where I^L is a diagonal matrix, in which I_{jj} is set to be 1 if x_j is originally labeled, and 0 otherwise.

Differentiating $Q(f)$ with respect to f , we have:

$$\frac{dQ}{df} \Big|_{f=f^*} = 2 \times [(I - D^{-1/2} W D^{1/2}) f^* + \mu I^L (f^* - y)] = 0 \quad (6)$$

With simple deduction, we obtain:

$$\mathbf{f}^* = (\mathbf{I} - \alpha \mathbf{S} - \beta \mathbf{A}^L)^{-1} \beta \mathbf{I}^L \mathbf{y} \quad (7)$$

where $\mathbf{S} = \mathbf{D}^{-1/2} \mathbf{W} \mathbf{D}^{1/2}$, $\mathbf{A}^L = \mathbf{I} - \mathbf{I}^L$, $\alpha = 1/(1 + \mu)$ and $\beta = \mu/(1 + \mu)$.

When the number of data is large, we can replace it with an iteration process:

$$\mathbf{f}(t+1) = (\alpha \mathbf{S} + \beta \mathbf{A}^L) \mathbf{f}(t) + \beta \mathbf{I}^L \mathbf{y} \quad (8)$$

When the iterative process converges, we obtain the modified classification score vector \mathbf{f}^* . Based on above regularization framework, our regularized semi-supervised LDA algorithm (r-SSLDA) is summarized as Algorithm 1. In our r-SSLDA, the affinity matrix \mathbf{W} is given by

$$\mathbf{W} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2) \quad (9)$$

where $\sigma = 5$.

Algorithm 1. Regularized Semi-supervised LDA algorithm

Input: (1) A training image set $\mathcal{X} = \mathcal{X}^L \cup \mathcal{X}^U = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \subset \mathbb{R}^d$, where \mathcal{X}^L is a labeled image set, and \mathcal{X}^U is an unlabeled image set;
(2) Initialize the label vector $\mathbf{y} = (y_1, y_2, \dots, y_n)$, in which y_i is the label of image \mathbf{x}_i .

Output: The parameter of a supervised LDA model

Steps:

- (1) Take all the labeled images as a subset and train it with supervised LDA model. Use this model to predict the initial classification score \mathbf{f} of the entire dataset.
 - (2) Form the affinity matrix \mathbf{W} defined by equation (9).
 - (3) Solve equation (4) according to equation (8), and obtain the final final classification score \mathbf{f}^* , which indicates accurate labels of all the training images including both the labeled and unlabeled images.
 - (4) Given \mathbf{f}^* of the entire training dataset, assign image label y_i to unlabeled image $\mathbf{x}_i \in \mathcal{X}^U$ according to the sign of f_i^* .
 - (5) Train image classifiers using sLDA model and all the training images (all images are labeled), and the parameter of the sLDA model is the final classifier for each image category.
-

To directly encode class labels into r-SSLDA, we adopt supervised LDA [9, 10] as our topic model to training image classifiers. The main idea of r-SSLDA is the following. We first use initially labeled images and supervised LDA to train original

image classifiers. After training image classifiers, each image x_i will be assigned a real value f_i to indicate its class label. Then we use the regularization framework to modify every f_i , which is the essential part of r-SSLDA. As a result, an image in the training set has its own f_i to indicate an accurate label. With proper class labels, we reuse sLDA to obtain the final image classifiers.

3 Experiments

3.1 Data Preparation and Feature Extraction

We performed our experiments on the Caltech 101 image dataset [18] and Caltech 256 image dataset [15]. Compared with Caltech 101, images in Caltech 256 contain more complex clutters. We chose five categories (leopard, motorbike, watch, airplane, face) from Caltech 101 and five categories (bathtub, billiard, binocular, gorilla, grape) from Caltech 256 to form our dataset. We select these categories because they have more images than other categories. Figure 2 shows examples of these images. To train binary classifiers, we select the background-google category in Caltech 101 as negative examples. In total, we have 11 categories with over 3600 images in our experiments. To keep authority, each experiment ran 8 times. The final results were the average of 8 runs.



Fig. 2. Sample images in our experiments: (a) images from Caltech 101, including airplane, face, leopard, watch, and motorbike; (b) images from Caltech 256, including bathtub, billiard, binocular, gorilla, and grape

From these images, we extracted key points and their SIFT descriptors, and used k-means algorithm to quantize these SIFT descriptors into visual words [8, 16]. In the end, we generated 300 visual words to form our visual codebook. Each image was represented by the popular bag of visual words model.

3.2 Regularized Semi-supervised LDA vs. Fully Supervised LDA

We compared our proposed regularized semi-supervised Latent Dirichlet Allocation algorithm (r-SSLDA) with classic LDA, and fully supervised LDA (sLDA). For each category, we randomly selected 100 images as training data, and 100 images from the background category. In our settings for r-SSLDA, only 20% of the training images

were labeled. That is, 40 out of 200 training images were randomly selected and labeled, 20 images from given category as positive samples and 20 images from background category as negative samples. The residual 160 training images were unlabeled when training. We trained three visual concept classifiers, respectively, using each of the three methods. For each classifier, we performed binary classification on 200 test images (100 images from the correspondence category and 100 images from the background category). For all the methods, the topic number is set to 30. The classification performance was shown in Figure 3.

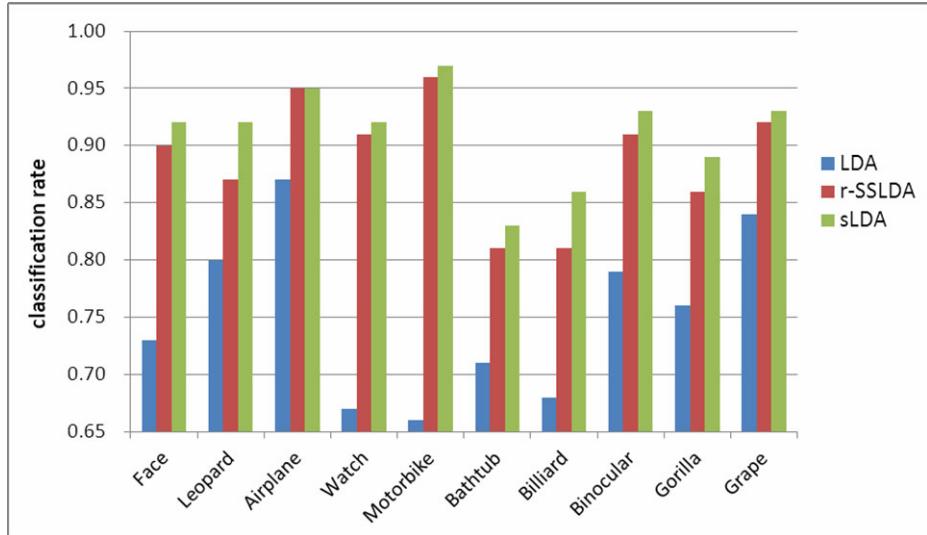


Fig. 3. Recognition results on the ten categories. The percentage of labeled images is 20%. The topic number is 30 for all the methods.

As seen from Figure 3, r-SSLDA outperforms classic LDA for all the 10 classes. It proves that supervised information is very important to improve the performance of the topic model. Compared with sLDA, though only 20% of the training images are labeled, our r-SSLDA incurred little loss on the classification rate, while significantly reducing the required labeled data. In practice, labeled images are very costly to obtain, while unlabeled images are readily available from the Internet. This makes our r-SSLDA more suitable and advantageous for real applications.

3.3 Regularized Semi-supervised LDA vs. Simple Semi-supervised LDA

There are many strategies to learn a topic model in a semi-supervised way. One of the simple strategies is to implement topic models twice. First, we use labeled images to train an initial classifier with sLDA. Then, we use the initial classifier to predict the label of unlabeled training data. After obtaining all the labels for all the training images, we use sLDA to train the visual concept classifier. We call this strategy simple Semi-supervised LDA (s-SSLDA). s-SSLDA is vulnerable to prediction errors

because of data noise and model bias. To reduce the prediction errors, our r-SSLDA refines the predictions using a regularization framework that simultaneously considers smoothness and consistency. To validate the efficiency of our regularization

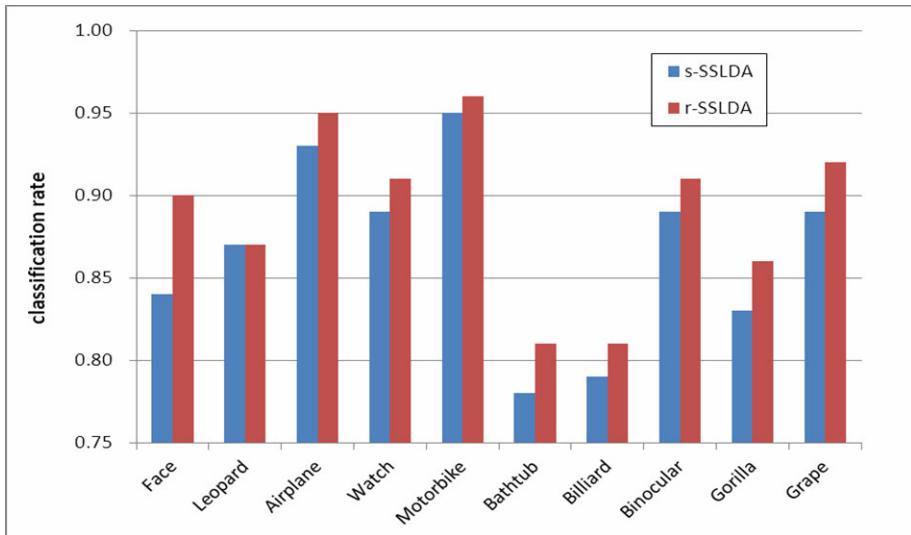


Fig. 4. Performance comparison between r-SSLDA and s-SSLDA across all the ten categories. 20% of the training images were labeled. The Topic number is 30.

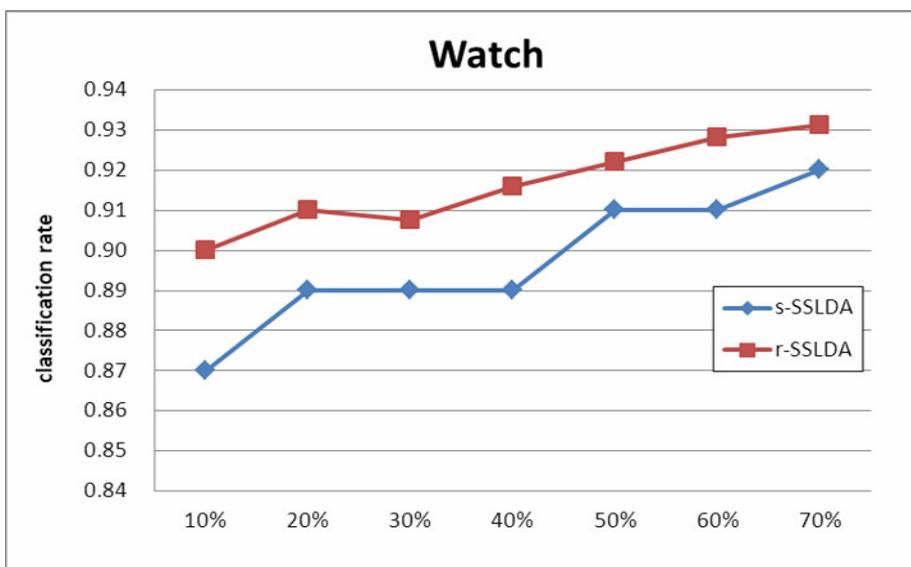


Fig. 5. Performance comparison between r-SSLDA and s-SSLDA over watch category with different percentages of labeled images, ranging from 10% to 70%. The topic number is 30.

framework, we compared r-SSLDA with s-SSLDA in all ten categories. Figure 4 showed the performance comparison between r-SSLDA and s-SSLDA, when the percentage of labeled training images is 20%. As we can see, our r-SSLDA outperformed s-SSLDA in most cases. This proves that our regularization framework is more efficient than the simple semi-supervised strategy for combining supervised and unsupervised information.

Furthermore, we changed the percentage of labeled images and compared the performance of r-SSLDA with that of s-SSLDA. Take the watch category for example, we changed the labeled percentage ranging from 10% to 70%. As shown in Figure 5, our r-SSLDA outperformed s-SSLDA over all the cases (percentages). Similar results were achieved in our experiments on face, airplane, motorbike, bathtub, billiard, gorilla, and grape. These experiments once again proved that, with the same number of labeled images, our regularization framework was more effective than s-SSLDA for combining a limited number of labeled training images with a large amount of unlabeled training images.

4 Conclusions

In this work, to take advantage of both limited labeled image data and a large amount of unlabeled image data, we developed a novel regularization framework for semi-supervised topic model learning. Based on this framework, we proposed a regularized Semi-Supervised Latent Dirichlet Allocation (r-SSLDA) for visual concept classifier learning. Unlike the solution in [12], we considered both semi-supervised properties and topic models simultaneously in the regularization framework. Experiments on Caltech 101 and Caltech 256 have shown that our r-SSLDA outperformed classic unsupervised LDA and achieved competitive performance against fully supervised LDA (sLDA), while drastically reducing the requirement of labeled training images. Moreover, experiments showed that our r-SSLDA was more effective than s-SSLDA in utilizing both unlabeled images and labeled images.

Furthermore, similar to most supervised learning algorithms, sLDA tends to overfit with the increase of labeled images. When the amount of labeled images is large, r-SSLDA is more stable than sLDA. In the future, we plan to evaluate our r-SSLDA algorithm on large scale datasets.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under contract No.60933013, National High Technology Research and Development Program of China (863) under contract No.2010ZX03004-003, the Fundamental Research Funds for the Central Universities, and the Science Research Fund of University of Science and Technology of China for Young Scholars.

References

- Chen, Y., Wang, J.Z.: Image categorization by learning and reasoning with regions. *JMLR* 5, 913–939 (2004)
- Blei, D.M., Jordan, M.I.: Modeling annotated data. In: *SIGIR* (2003)

3. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via pLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) *ECCV 2006*. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)
4. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their location in images. In: *IEEE International Conference on Computer Vision, ICCV 2005*, vol. 1, pp. 370–377 (2005)
5. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning Object Categories from Google’s Image Search. In: *IEEE International Conference on Computer Vision, ICCV 2005*, vol. 2, pp. 1816–1823 (2005)
6. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. *JMLR* 3, 993–1022 (2003)
7. Wang, Y., Mori, G.: Human Action Recognition by Semi-Latent Topic Models. *IEEE Trans. on Pattern Analysis and Machine Intelligence Special Issue on Probabilistic Graphical Models in Computer Vision T-PAMI* 31(10), 1762–1774 (2009)
8. Lowe, D.: Object recognition from local scale-invariant features. In: *IEEE International Conference on Computer Vision, ICCV 1999*, vol. 2, pp. 1150–1157 (1999)
9. Blei, D., McAuliffe, J.: Supervised topic models. In: *Advances in Neural Information Processing Systems*, vol. 21 (2007)
10. Wang, C., Blei, D., Fei-Fei, L.: Simultaneous image classification and annotation. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, June 20-25 (2009)
11. Zhu, X.: Semi-Supervised Learning Literature Survey. *Computer Sciences Technical Report* 1530, University of Wisconsin-Madison (July 19, 2008)
12. Zhuang, L., She, L., Jiang, Y., Tang, K., Yu, N.: Image Classification via Semi-supervised pLSA. In: *Proceedings of the 2009 Fifth International Conference on Image and Graphics (ICIG 2009)*, September 20-23, pp. 205–208 (2009)
13. Wang, C., Zhang, L., Zhang, H.-J.: Graph-based Multiple-Instance Learning for Object-based Image Retrieval. In: *Proceeding of the 1st ACM International Conference on Multimedia Information Retrieval*, October 30-31, pp. 156–163 (2008)
14. Zhou, D., Bousquet, O., Lal, N.T., et al.: Learning with local and global consistency. In: *NIPS* (2003)
15. Caltech-256 Dataset, <http://authors.library.caltech.edu/7694/>
16. Kadir, T., Brady, M.: Saliency, scale and image description. *International Journal of Computer Vision* 45(2), 83–105 (2001)
17. Wang, Y., Mori, G.: Human Action Recognition by Semilatent Topic Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(10), 1762–1774 (2009)
18. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: *Proceeding of IEEE Computer Vision and Pattern Recognition 2004, Workshop on Generative-Model Based Vision* (2004)

Boosted Scene Categorization Approach by Adjusting Inner Structures and Outer Weights of Weak Classifiers

Xueming Qian, Zhe Yan, and Kaiyu Hang

Department of Information and Communication Engineering,
Xi'an Jiaotong University
710049 Xi'an, China
qianxm@mail.xjtu.edu.cn

Abstract. Scene categorization plays an important role in computer vision and image content understanding. It is a multi-class pattern classification problem. Usually, multi-class pattern classification can be completed by using several component classifiers. Each component classifier carries out discrimination of some patterns from the others. Due to the biases of training data, and local optimal of weak classifiers, some weak classifiers may not be well trained. Usually, some component classifiers of a weak classifier may be not act well as the others. This will make the performances of the weak classifier not as good as it should be. In this paper, the inner structures of weak classifiers are adjusted before their outer weights determination. Experimental results on three AdaBoost algorithms show the effectiveness of the proposed approach.

Keywords: Pattern Classification, Back-propagation Networks, AdaBoost.

1 Introduction

Multi-class pattern classification covers a wide range of applications. Most approaches for multi-class pattern classification can be decomposed into multiple binary pattern classification problems [1]. One is using one-versus-all (OVA) classifiers based approach. Another is using the one-versus-one (OVO) classifiers based approach. The other is using p -versus- q ($1 \leq p, q \leq N - 1$; N is class number) classifiers based approach [1]. The OVA approach builds a binary classifier to distinguish every class from the rest of classes. The OVO approach builds a binary classifier for every pair of two classes. For an N -class ($N > 2$) pattern classification problem, N OVA classifiers or $N^*(N-1)/2$ OVO classifiers are needed to carry out exact classification [1].

Usually, it is hard to train a very robust classifier [2]-[6]. It is very easy to train dozens of weak classifiers with their performances are just good than random guessing. How to fuse dozens of weak classifiers to be a strong one is far more important than to train a single strong classifier [7],[8]. AdaBoost algorithms can fulfill this by assigning appropriate weights for weak classifiers. AdaBoost algorithms turn out to be very effective in machine learning and pattern recognition. In AdaBoost algorithms, classifiers are called weak classifiers because they are not expected to be

perfect. Many of them are just better than random guessing. Weak classifiers can be fused to be strong classifiers by setting weights according to their performances [7]. Usually, the weak classifiers with high performances have large weights while the poor classifiers have small weights [7, 8]. AdaBoost algorithms can be implemented either online or offline [8]. The online AdaBoost algorithms train weak classifiers and determine their weights in a unified framework using the same training set [7]. The online AdaBoost algorithms call weak learner to get a weak classifier with respect to the distribution of weighted training samples. The offline AdaBoost algorithms can be viewed as extended versions of features selection problem where weak classifiers are trained before their weights determination [8]. In the real world, some weak classifiers should be well trained except that some component classifiers are not as good as expected. If inner structures of weak classifiers can be adjusted and the influences of the inferior component classifiers can be decreased. However the offline AdaBoost algorithms do not pay attention to this problem. It is out scope of the offline AdaBoost algorithms and they relay on the weak learners and doing nothing to remedy inferior weak classifiers. In this paper, the inner structures of weak classifier are adjusted before assigning their outer weights. Thus the inferior weak classifiers have some chances to be adjusted to be performed better than their originals.

Scene categorization is a specified application of multi-class pattern categorization. Bag-of-Words (BoW) based approaches model objects in a scene/image as geometric-free structures [2]-[6], [8]-[16]. In [16], BoW based approaches represent objects with rigorous geometric structures by modeling the relationships of different parts. Usually, the local patches of an image are assumed to be independent from each other [10]. This assumption simplifies the computations for the ignorance of the spatial co-occurrences and dependences of local patches. While, in some applications, the co-occurrences, dependences and linkages of the salient parts of images are also modeled to improve scene categorization performances. These methods aim at training robust classifiers to fulfill scene categorization. Despite of constructing robust models, effective feature representation is also paid much attention [16], [3]-[6].

This paper is based on our previous work [8], where BP networks are served as the basic weak classifiers. Compared to our previous work [8], main contributions of this paper are weak classifiers inner structures adjusting and outer weights learning. The rest of this paper is organized as follows: In Section 2 AdaBoost algorithms are briefly overviewed. In Section 3 the proposed scene categorization approach is illustrated in detail. In Section 4 experiments are given. Finally, conclusions are drawn in Section 5.

2 Overview AdaBoost Algorithms

The diagram of AdaBoost algorithms for multiple class pattern classification is shown in Figure 1. AdaBoost algorithms determine the outer weights α_m ($k = 1, \dots, M$) of weak classifiers according to their performances on some weight training samples. Each weak classifier (denoted WeakC_m) consists of N component classifiers (denoted CompC_n). In this paper, each component classifier CompC_n ($n = 1, \dots, N$) is a one-versus-all back-propagation network or support vector machine which carrying out discrimination for class n and the other $N-1$ classes.

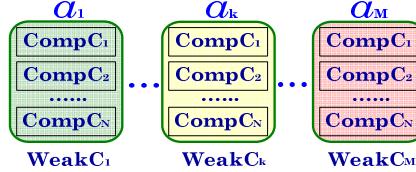


Fig. 1. Diagram of multi-class AdaBoost algorithms by assigning outer weights α_m for each of weak classifier. N and M are the numbers of class and weak classifiers.

AdaBoost.M1, AdaBoost.M2 and AdaBoost.MT are three AdaBoost algorithms for multi-class pattern classification [8]. They determine the weights of pre-trained M weak classifiers by a set of weight training samples. In AdaBoost.M1, each weak classifier outputs a hard decision for the test sample x . The final output $H(x)$ is related to weighted votes of all weak classifiers.

$$H(x) = \arg \max_{y \in Y} \sum_{t=1}^M \alpha_t [\![h_t(x) = y]\!]; Y = \{1, \dots, N\} \quad (1)$$

where α_t is the weight of weak classifier h_t , and

$$[\![h_t(x) = y]\!] = \begin{cases} 1 & h_t(x) = y \\ 0 & h_t(x) \neq y \end{cases}; y \in Y = \{1, \dots, N\} \quad (2)$$

where $h_t(x)$ outputs a hard label

$$h_t(x) = \arg \max_{y \in Y} \{h_t(x, y)\}; Y = \{1, \dots, N\} \quad (3)$$

where $h_t(x, y)$ is the response of weak classifier h_t to the label y .

During weak classifiers' weights learning, AdaBoost.M1 views a weak classifier useless when the correct recognition rate is small than or equals to 1/2. Generally, this requirement is too serious for weak classifiers. In order to make weak classifiers contributive to multi-class pattern classification, AdaBoost.M2 and AdaBoost.MT make full use of the response of each component classifier of a weak classifier to make decision. Different from AdaBoost.M1, in AdaBoost.M2 each weak classifier outputs a response vector rather than a hard decision [7]. Each component of the response vector is in the range [0, 1]. Labels with large values are considered to be believable. Labels with small values are considered unbelievable. For a given test sample x , AdaBoost.M2 and AdaBoost.MT [8] output the estimated label as follows

$$H(x) = \arg \max_{y \in Y} \sum_{t=1}^T \alpha_t h_t(x, y) \quad (4)$$

where $\alpha_t h_t(x, y)$ is the weighted response of weak classifier h_t to the label y .

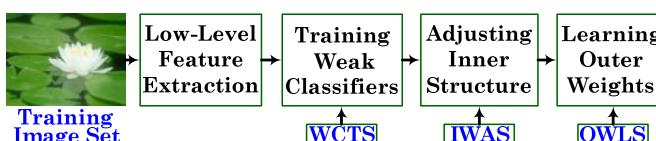


Fig. 2. Flowchart of the Training Process of the Proposed Scene Categorization Approach

3 Boosted Scene Categorization by Adjusting Inner Structures and Determining Outer Weights of Weak Classifiers

The category of an input image is determined using the proposed approach as follows:

- 1) Extract low-level features; four low-level features are extracted for the input image.
- 2) Get the response vector of each weak classifier according to the input feature;
- 3) Adjust the response vector of each weak classifier according to the optimized inner weights and biases;
- 4) Output the estimated label by the weighted responses of each weak classifier.

Before carrying out scene classification, we need to train dozens of weak classifiers, adjust their inner structures and determine their outer weights. The above three steps are the core of the training process which is shown in Figure 2. The training process consists of following four steps: 1) low-level feature extraction; 2) training weak classifiers; 3) adjusting inner structures; and 4) learning outer weights. Three training sets are utilized. They are weak classifiers' training set (WCTS), inner structure adjusting set (IWAS) and weak classifiers' outer weights learning set (OWLS) respectively. Compared to our previous work [8], we need IWAS to adjust inner structures of weak classifiers.

3.1 Low-Level Feature Extraction

Four features: SPM [4], PHOG [3], GIST [5] and HWVP [6] are served as the input features to train weak classifiers and to carry out scene categorization. The four descriptors have been shown their effectiveness in feature representation [3]-[6],[8]. Actually, from scene categorization points of view, more weak classifiers trained using various complementary features better results can be achieved. Now we give a brief overview of the four features.

SPM feature is a local scale invariant descriptor by using spatial pyramid transforms. In the extraction of **SPM** [4], the 128 dimensional SIFT features [16] are converted into visual words. SPM is composed of visual words histograms of an image at various spatial pyramids. SPM is robust to the variations of rotation, scaling and illumination [4]. **PHOG** represents an image with histograms of orientation gradients over spatial pyramids. Each bin in PHOG represents the number of edges that have orientations within a certain angular range. It has the advantages to represent image with certain global and local shape information. **HWVP** represents texture information of an image using hierarchical wavelet packet transform [7]. HWVP improves the discrimination power for images by utilizing sub-bands filtering in hierarchical wavelet packet domain [7]. **GIST** is an effective high dimensional texture descriptor. This feature takes advantages of Gabor transform. It is robust to represent image texture information with oriented multiple-scale based filtering [6].

In following part of this paper, let \mathbf{x} denote the input feature of an image. $\mathbf{x} = (x_1, \dots, x_d) \in \mathbb{R}^d$ is a d dimensional vector. The dimension d is a feature related parameter. The dimensions of SPM, PHOG, HWVP, GIST are $d=6300, 850, 850, 640$ respectively. For more details, please turn to [8].

3.2 Training Weak Classifiers

For the N -class scene categorization problem, N one-versus-all component classifiers are combined to determine the exact label [1],[8]. During training, b samples per category are randomly selected and served as the weak classifiers' training set (i.e. WCTS). V weak classifiers for each of the four features are trained by running the training process V times. In this paper, we set $V=20$. In each training process, b samples per category are randomly selected from WCTS to train a weak classifier. For the training of the k -th CompC of a WeakC, the b training images from the k -th class are selected as positive samples and $b^*(N-1)$ images from the other $N-1$ classes are selected as negative samples. Some of the samples may be utilized more than once and some of them may not be selected in each process, thus the V WeakCs should have some complementary in the fusion stage.

3.3 Boosted Scene Categorization Approach by Adjusting Inner Structure and Outer Weights of Weak Classifiers

Let $\mathbf{R}_q = \langle R_q^1, \dots, R_q^N \rangle$ denote the response vector of the q -th weak classifier and R_q^k denote the response of the k -th ($k=1, \dots, N$) component classifier of the q -th ($q=1, \dots, M$) weak classifier. Before fusing the M weak classifiers, the inner structure of a weak classifier is further adjusted by finding optimal biases vector $\boldsymbol{\eta}_q = \langle \eta_q^1, \dots, \eta_q^N \rangle$ and inner weight vector $\boldsymbol{\beta}_q = \langle \beta_q^1, \dots, \beta_q^N \rangle$. After adjusting inner structures of weak classifier, scene categorization is carried out by fusing the adjusted weak classifiers using existing AdaBoost algorithms.

Correspondingly, the boosted approach of AdaBoost.M2, and AdaBoost.MT after weak classifiers' inner structures adjusting is as follows

$$H(\mathbf{x}) = \arg \max_{y \in Y} \sum_{i=1}^M \alpha_i \times h_i(\mathbf{x}, y) = \arg \max_{y \in Y} \sum_{i=1}^M \alpha_i \times \left[\beta_i^y \left(R_i^y(\mathbf{x}) - \eta_i^y \right) \right]; Y = \{1, \dots, N\} \quad (5)$$

where $h_i(\mathbf{x}, y)$ is the t -th ($t=1, \dots, M$) adjusted classifier. And for the AdaBoost.M1, the boosted approach after inner structure adjusting is as follows:

$$H(\mathbf{x}) = \arg \max_{y \in Y} \sum_{t=1}^M \alpha_t \llbracket h_t(\mathbf{x}, y) \rrbracket; Y = \{1, \dots, N\} \quad (6)$$

where $\llbracket h_t(\mathbf{x}, y) \rrbracket$ is determined as follows

$$\llbracket h_t(\mathbf{x}, y) \rrbracket = \begin{cases} 1 & \beta_t^y \left(R_t^y(\mathbf{x}) - \eta_t^y \right) = \max_{k \in Y} \left\{ \beta_t^k \left(R_t^k(\mathbf{x}) - \eta_t^k \right) \right\} \\ 0 & \beta_t^y \left(R_t^y(\mathbf{x}) - \eta_t^y \right) \neq \max_{k \in Y} \left\{ \beta_t^k \left(R_t^k(\mathbf{x}) - \eta_t^k \right) \right\} \end{cases}; \quad Y = \{1, \dots, N\}. \quad (7)$$

The proposed algorithm can be simplified into the original algorithms if $\boldsymbol{\beta}_q = \mathbf{1}$ and $\boldsymbol{\eta}_q = \mathbf{0}$ ($\mathbf{1}$ and $\mathbf{0}$ are vectors with dimensions $1 \times N$, each elements in them are with sample values 1 and 0.). In order to adjust the inner structure of each weak classifier, S samples $\{(\mathbf{x}^1, y^1), \dots, (\mathbf{x}^S, y^S)\}$ with their labels $y^i \in Y = \{1, \dots, N\}$ are utilized to find the

optimal parameters $\boldsymbol{\theta}_q^* = \langle \boldsymbol{\eta}_q^*, \boldsymbol{\beta}_q^* \rangle$ of the adjusted weak classifiers. Correspondingly, the objective function is related to the correct recognition rate which is expressed as follows

$$f(\boldsymbol{\theta}_q) = \frac{1}{S} \sum_{s=1}^S \left[h_q(\mathbf{x}^s, y^s) \right]; \quad q = 1, \dots, M \quad (8)$$

where $\left[h_q(\mathbf{x}^s, y^s) \right]$ is a function of $\boldsymbol{\beta}_q$ and $\boldsymbol{\eta}_q$ as shown in Eq.(7).

The aim of inner structure adjusting is to find optimal parameters $\boldsymbol{\theta}_q^* = \langle \boldsymbol{\eta}_q^*, \boldsymbol{\beta}_q^* \rangle$ by maximizing the correct recognition rate

$$\boldsymbol{\theta}_q^* = \langle \boldsymbol{\eta}_q^*, \boldsymbol{\beta}_q^* \rangle = \arg \max_{\boldsymbol{\theta}_q} f(\boldsymbol{\theta}_q); \quad q = 1, \dots, M \quad (9)$$

In order to sure the adjusting is valid; the following constraints must be satisfied

$$f(\boldsymbol{\theta}_q^*) \geq f(\boldsymbol{\theta}_q^0); \quad q = 1, \dots, M \quad (10)$$

where $\boldsymbol{\theta}_q^0$ denote the initial parameters of weak classifier without adjusting its inner structure, i.e.

$$\boldsymbol{\theta}_q^0 = \langle \boldsymbol{\eta}_q^0, \boldsymbol{\beta}_q^0 \rangle; \quad \boldsymbol{\eta}_q^0 = \mathbf{0}; \quad \boldsymbol{\beta}_q^0 = \mathbf{1}; \quad q = 1, \dots, M \quad (11)$$

3.4 Genetic Algorithm Based Parameters Optimization

Genetic algorithm (GA) is utilized to find optimal parameter vector $\boldsymbol{\theta}_q^* = \langle \boldsymbol{\eta}_q^*, \boldsymbol{\beta}_q^* \rangle$ ($q = 1, \dots, M$) as shown in Eq.(9). In this paper, we assume that only some component classifiers of a weak classifier are not performed very well. The inner structures of each weak classifier only needed to be slightly adjusted. So, the inner weights are assigned in the range $\beta_q^k \in [0.8, 1.2]$ ($q = 1, \dots, M; k = 1, \dots, N$). The biases are in the range $\eta_q^k \in [-0.2, 0.2]$. Actually, the inner weights and biases vectors can be set to be the values in large range. However, this will increase the computation cost. The flowchart of GA based optimization is as follows:

- 1) Initialize parameters of GA: crossover probability $Pc=0.8$, mutation probability $Pm=0.05$, population size $Ps=400$, maximum iteration times $Im=50000$, minimum error variation $Em=10^{-6}$, initial evolution step $gen=1$.
- 2) Generate Ps individuals and encode them into chromosomes;
- 3) Calculate fitness values of each chromosome according to Eq.(8); The individuals with large fitness values correspond to the parameters with high correct recognition rates.
- 4) Update evaluation generations $gen=gen+1$; Select Ps chromosomes to the next generation according to the fitness. The selection probability of each chromosome is calculated as follows

$$P(k) = fit(k) / \sum_{k=1}^{Ps} fit(k) \quad (12)$$

- 5) Generate Ps new individuals by genetic operations (crossover and mutation) according to the crossover probability Pc and mutation probability Pm ;

- 6) Repeat step 3) to step5) if the evaluation step gen less than Im and the performance improvement of neighboring two generations is larger than Em ;
 7) Select the chromosomes with highest fitness as the final output $\boldsymbol{\theta}_q^*$.

For M weak classifiers, each classifier is adjusted separately. In order to make sure the final scene categorization can benefit from the adjusted weak classifiers, adjusting validation is needed. If $f(\boldsymbol{\theta}_q^*) \geq f(\boldsymbol{\theta}_q^0)$ is not satisfied, then the original weak classifier is utilized in the final fusion stage.

4 Experimental Results and Discussions

Experiments are conducted to evaluate the performance of boosted scene categorization approach by adjusting the inner structures and determining outer weights of weak classifiers. The experiments are on OT [5], and Sport Event [18] datasets.

Comparisons of the original AdaBoost algorithms (AdaBoost.M1, AdaBoost.M2 and AdaBoost.MT) and the proposed approach are made. Back-propagation networks are utilized as the basic weak classifiers. In the experiments, accurate recognition rate (AR) [8] is utilized to evaluate scene categorization performance.

In Table 1 and Table 2, scene categorization performances by utilizing totally T training samples per category are shown. The numbers of samples for weak classifier training, inner structure adjusting, and AdaBoost algorithms are b , g , and a respectively. The average accurate recognition rates and their deviations of 10 times run of the proposed AdaBoost algorithms are shown. $g=0$ corresponds to the original AdaBoost algorithms and $g \neq 0$ denotes the proposed approach.

4.1 Experimental Results on OT Dataset

Table 1 shows scene categorization performances for this dataset using 100, 80, 50 and 30 training samples per category and all the remaining samples are utilized for performance evaluation. In the circumstances that 100 samples per category (i.e. $T=100$) are utilized during training with $b=50$ and $a+g=50$, the performances of several combinations of a and g are shown in Table 1. When all the 50 samples are utilized by AdaBoost algorithms (i.e. $a=50$) to determine outer weights and no inner structure adjusting (i.e. $g=0$) is adopted, scene categorization performances of AdaBoost.M1, AdaBoost.M2 and AdaBoost.MT are 75.44%, 82.16% and 82.60% respectively. When about half of the weight training samples per category (i.e. $g=10, 15, 20$) are selected for adjusting inner structures, the average scene categorization performances of AdaBoost.M1, AdaBoost.M2 and AdaBoost.MT are improved by about 0.98%, 0.47% and 1.1% respectively over the original AdaBoost algorithms.

In the circumstance that totally 50 samples per category are selected as training set (i.e. $T=50$) with $b=40$ and $a+g=10$, the average performances of AdaBoost.M1, AdaBoost.M2 and AdaBoost.MT are improved by about 2.63%, 0.53% and 0.05% respectively. The average accurate recognition rates for the proposed approach are 69.55%, 74.54% and 74.78% respectively for the original AdaBoost.M1, AdaBoost.M2 and AdaBoost.MT algorithms under $b=20$ and $a+g=10$.

Table 1. Scene categorization performances for OT dataset under various testing conditions. Totally, T training samples per category are utilized for weak classifiers training, inner structure adjusting and outer weights learning. The samples numbers of weak classifiers training, inner structure adjusting, and weight learning are b , a and g respectively.

T	b	a	g	AdaBoost.M1	AdaBoost.M2	AdaBoost.MT
100	50	50	0	75.44	82.16	82.60
		30	20	76.67±0.43	82.34±0.17	83.69±0.51
		20	30	76.58±0.39	82.46±0.29	82.99±0.14
		25	25	76.01±0.18	83.20±0.22	84.14±0.62
80	50	30	0	76.01	83.20	83.39
		20	10	78.07±0.56	83.72±0.13	84.55±0.32
		15	15	77.55±0.34	83.81±0.23	84.56±0.29
		10	20	77.39±0.36	83.56±0.19	83.78±0.11
50	40	10	0	75.29	80.94	79.98
		6	4	77.57±0.53	81.29±0.16	80.39±0.13
		5	5	78.14±0.86	81.08±0.14	80.56±0.18
		4	6	78.06±0.55	81.21±0.14	80.12±0.27
30	20	10	0	65.72	73.99	74.50
		8	2	70.50±0.63	74.45±0.21	74.81±0.19
		7	3	70.27±0.51	74.46±0.16	74.77±0.15
		6	4	70.14±0.45	74.62±0.14	75.06±0.36
		5	5	69.79±0.97	74.56±0.15	74.68±0.16
		4	6	68.53±0.58	74.49±0.18	74.75±0.19
		3	7	68.24±0.75	74.55±0.14	74.67±0.13
		2	8	69.36±0.63	74.67±0.16	74.73±0.20

4.2 Experimental Results on Sport Event Dataset

Table 2 shows scene categorization performances for Sport Event dataset using 20 and 50 training samples per category and all the remaining samples are utilized for performance evaluation. When all the 20 samples per category (i.e. $T=20$) are utilized for training weak classifiers with $b=15$, $a=5$ and $g=0$, the corresponding categorization performance of AdaBoost.M1, AdaBoost.M2 and AdaBoost.MT are 50.9%, 64.29% and 64.69% respectively. When about half of the weight training samples per category (i.e. $a=3$, $g=2$; $a=2$, $g=3$) are selected for adjusting the structures of weak classifiers, the average improvements of scene categorization performances are 0.39%, 0.37% and 1.57% respectively for AdaBoost.M1, AdaBoost.M2 and AdaBoost.MT. For the cases when 50 samples per category are utilized (i.e. $T=50$),

with $b=40$ and $a+g=10$, the proposed approach improves scene categorization performances by about 2.76%, 0.48% and 0.97% in average over AdaBoost.M1, AdaBoost.M2 and AdaBoost.MT.

Boosted scene categorization approaches by fusing the adjusted inner structures of weak classifiers achieve better performances. It is also clear that the weak classifiers' training is the most import part of this paper. When the totally training samples are equal, only the weak classifiers training sample number is sufficient, better performances can be achieved.

Table 2. Scene categorization performances for Sport Event dataset under various testing conditions. Totally, T training samples per category are utilized for weak classifiers training, inner structure adjusting and outer weights learning. The samples numbers of weak classifiers training, inner structure adjusting, and weight learning are b , a and g respectively.

T	b	a	g	AdaBoost.M1	AdaBoost.M2	AdaBoost.MT
20	15	5	0	50.90	64.29	64.69
		3	2	51.47±0.16	64.48±0.12	66.36±0.38
		2	3	51.10±0.13	64.84±0.21	66.16±0.35
50	40	10	0	57.76	77.22	74.55
		6	4	60.89±0.72	77.65±0.16	75.47±0.28
		5	5	60.02±0.53	77.46±0.18	75.34±0.13
		4	6	60.65±0.66	77.99±0.22	75.74±0.29

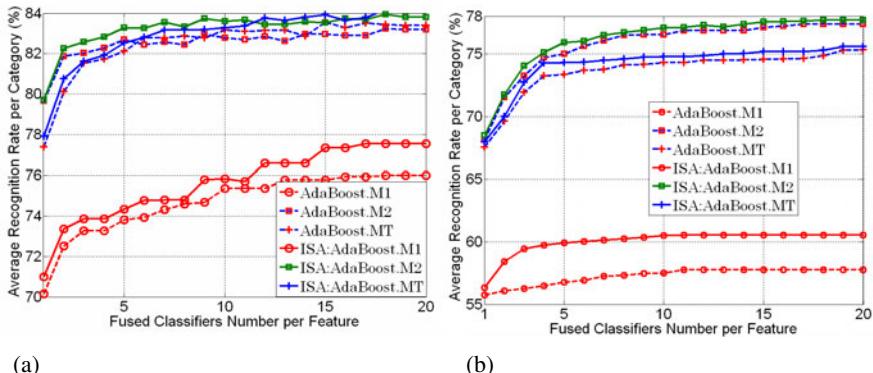


Fig. 3. Boosted Scene Categorization Performances for OT dataset and Sport Event datasets using Inner Structure Adjusting approach (a) OT dataset under $T=80$, $b=50$, $a=15$, $g=15$. (b) Sport Event under $T=50$, $b=40$, $a=5$, $g=5$.

4.3 Performances Versus Fused Weak Classifier Number

In Figure 3, boosted scene categorization performances versus fused weak classifier number per feature are shown. AdaBoost.M1, AdaBoost.M2 and AdaBoost.MT denote

the original AdaBoost Algorithms. And ISA:AdaBoost.M1, ISA:AdaBoost.M2 and ISA:AdaBoost.MT denote the AdaBoost Algorithms by fusing the weak classifiers after inner structure adjusting. Figure 3(a) shows the performance of $T=100, b=50, a=25$ and $g=25$. The corresponding boosted scene categorization performances versus fused weak classifier number per feature of AdaBoost.M1, AdaBoost.M2 and AdaBoost.MT (under $T=50, b=40, a=10$), and ISA:AdaBoost.M1, ISA:AdaBoost.M2 and ISA:AdaBoost.MT (under $T=50, b=40, a=5, g=5$) are shown in Figure 3 (b).

5 Conclusion

In this paper, a boosted scene categorization approach is proposed by modifying inner structures of weak classifiers and assigning them appropriate outer weights. AdaBoost algorithms are utilized to determine the weights of the weak classifiers after inner structure adjusting. Experimental results show that the proposed approach improves scene categorization performance. Adjusting the inner weights and biases of each weak classifier can improve scene categorization performances. The ratio of the training samples of weak classifier, structure modification and AdaBoost algorithms are important for the final performance. Only the weak classifiers are well trained, the inner structure adjusting and outer weights learning approach can further improve scene categorization performances. In AdaBoost algorithms the weak classifiers are not expected to be perfect, if the basics weak classifiers are strong enough then better performance can be achieved by fusing them.

Acknowledgments. This work is supported in part by the National Natural Science Foundation of China Project (NSFC, No.60903121).

References

1. Ou, G., Murphey, Y.: Multi-class pattern classification using neural networks. *Pattern Recognition* 40, 4–18 (2007)
2. Wu, L., Hu, Y., Li, M., Yu, N., Hua, X.: Scale-invariant visual language modeling for object categorization. *IEEE Trans. Multimedia* 11, 286–294 (2009)
3. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. In: CIVR (2007)
4. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. CVPR (2006)
5. Torralba, A., Willian, K., Freeman, T., Rubin, M.: Context-based vision system for place and object recognition. In: ICCV (2003)
6. Qian, X., Liu, G., Guo, D., Li, Z., Wang, Z., Wang, H.: Object categorization using hierarchical wavelet packet texture descriptors. In: ISM, pp. 44–51 (2009)
7. Freud, Y., Schapire, R.: A decision-theoretic generalization of online learning and an application to boosting. *Journal of Computer and System Sciences* 55, 119–139 (1997)
8. Qian, X., Yan, Z., Hang, K., Liu, G., Wang, H., Wang, Z., Li, Z.: Scene categorization using boosted back-propagation neural networks. In: Qiu, G., Lam, K.M., Kiya, H., Xue, X.-Y., Kuo, C.-C.J., Lew, M.S. (eds.) PCM 2010. LNCS, vol. 6297, pp. 215–226. Springer, Heidelberg (2010)

9. Monay, F., Gatica-Perez, D.: PLSA-based image auto-annotation:constraining the latent space. In: ACM Multimedia (2004)
10. Li, F., Perona, P.: A Bayesian hierarchy model for learning natural scene categories. In: CVPR (2005)
11. Zheng, Y., Zhao, M., Neo, S., Chua, T., Tian, Q.: Visual synset: towards a higher-level visual representation. In: CVPR (2008)
12. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. International Journal of Computer Vision (2007)
13. Bosch, A., Zisserman, A., Munoz, X.: Scene classification using a hybrid generative/discriminative approach. IEEE TPAMI 30, 712–727 (2008)
14. Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. In: CVPR (2008)
15. Cao, L., Li, F.: Spatially coherent latent topic model for concurrent object segmentation and classification. In: ICCV (2007)
16. Holub, A., Perona, P.: A discriminative framework for modeling object classes. In: ICCV (2005)
17. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
18. Li, L., Li, F.: What, where and who? Classifying events by scene and object recognition. In: ICCV (2007)

A User-Centric System for Home Movie Summarisation

Saman H. Cooray, Hyowon Lee, and Noel E. O'Connor

CLARITY: Centre for Censor Web Technologies, Dublin City University, Ireland
samam.hcooray@gmail.com

Abstract. In this paper we present a user-centric summarisation system that combines automatic visual-content analysis with user-interface design features as a practical method for home movie summarisation. The proposed summarisation system is designed in such a manner that the video segmentation results generated by the automatic content analysis tools are further subject to refinement through the use of an intuitive user-interface so that the automatically created summaries can be effectively tailored to each individual's personal need. To this end, we study a number of content analysis techniques to facilitate the efficient computation of video summaries, and more specifically emphasise the need for employing an efficient and robust optical flow field computation method for sub-shot segmentation in home movies. Due to the subjectivity of video summarisation and the inherent challenges associated with automatic content analysis, we propose novel user-interface design features as a means to enable the creation of meaningful home movie summaries in a simple manner. The main features of the proposed summarisation system include the ability to automatically create summaries of different visual comprehension, interactively defining the target length of the desired summary, easy and interactive viewing of the content in terms of a storyboard, and manual refinement of the boundaries of the automatically selected video segments in the summary.

Keywords: video summarisation; camera motion estimation; user interaction; home movie.

1 Introduction

We are witnessing increasing numbers of video repositories being created by many home users due mainly to the ubiquitous nature of video capture devices that are becoming more affordable. One significant requirement in dealing with the management of such raw video archives is automated video summarisation; a task that can effectively help users organise many hours of movie material, leading to improved viewing experience overall.

Video summarisation, in general, refers to the mechanism of creating a concise version of the original digital video, to facilitate users to perform efficient browsing, search and retrieval of multimedia content in large-scale video archives. In a task like browsing of large video archives, a video summary can provide users

with useful information to get a rough idea about the original content of the video in a much shorter time. For summarisation, home movies pose a particularly difficult challenge due to unrestricted capture and lack of storyline present in the content. To create compact movie clips from an unedited video content, home users currently rely on some existing video editing tools, such as Apple's iMovie¹, and Adobe Premiere². Unfortunately, using any of these tools is a laborious and cumbersome process. An automated solution must also ensure that the following principles are satisfied, in order that a summary meets the purpose of real users.

- The most important events and activities are included in the summary.
- Redundant events are sufficiently filtered and excluded from the summary.

Addressing numerous challenges in home movie summarisation, a substantial amount of work has been reported in the literature, and in particular, the last decade has seen greater interest from the research community towards the development of summarisation tools [1]2[3]4[5]6[7]8[9]. An automatic home video abstraction method was proposed by Lienhart [1] using a date/time based clustering approach and a shot shortening method based on sound features. Kender and Yeo [2] presented a home-video summarisation system based on the use of a zoom-and-hold filter as an implicit human visual attention rule applied when capturing home movies. A probabilistic hierarchical clustering approach was proposed by the authors of [3] using visual and temporal features to discover cluster structure in home video. Huang *et al.* [4] presented an intelligent home video management system using fast-pan elimination, face-shot detection, etc. Recently, Mei *et al.* [5] proposed a novel home video summarisation method based on the exploitation of the user's intention at the time of capture, as a complementary mechanism to existing content analysis schemes. The authors of [6] make use of the home users' photo libraries to infer their preferences for video summarisation. Wang *et al.* [7] presented an information-theoretic approach to content selection as an effective method for selecting the most important content in home video editing. By modeling the co-occurrence statistics between characters (who) and scenes (where), the authors create a compact representation of raw footage from which they extract the most important content using a joint entropy measure. More recently, we presented an interactive and multi-level framework for home video summarisation, combining automatic content analysis with user interaction to create visually comprehensive summaries [8]. Peng *et al.* [9] proposed a user experience model for home video summarisation, taking into account the user's reaction such as eye movement and facial expression when viewing videos. Despite growing interest from the research community, automatic summarisation of home movie remains a challenging research topic due to the presence of unstructured storyline and unrestricted capture in home video footage.

Recently, there has also been a significant body of work on user-interface technologies for home video editing and authoring. The Hitchcock system [10]

¹ <http://www.apple.com/ilife/imovie/>

² <http://www.adobe.com/products/premiereel/>

performs automatic motion analysis of the raw video to determine which parts of the video, i.e. clips, should be included in the summary. Users can interactively override the start and end time of a clip by re-sizing its keyframe. Campanella *et al.* [11] developed the Edit While Watching (EWW) system that has the ability to automatically create an edited version of the raw home video and allow the user to refine results interactively. Upon loading a raw video into the system, a set of short video segments are created based on the analysis of low-level features, such as camera motion, contrast and luminosity, which collectively form the automatically edited version of the video. The system then allows the user to add/remove content to/from the edited version at sub-shot, shot or scene level. Based on the study of existing research and commercial frameworks, it is clear that a practical approach to home movie summarisation should consider a user-centric scenario, ensuring that the work on the part of the user is reduced to the best possible level. Furthermore, user studies carried out by the authors of [12][13] suggest that home users always want to have the flexibility to tailor the automatically created summaries according to their personal needs.

In this paper, we present a home movie summarisation system, focussing on design concepts of the user-interface while taking into account the challenges associated with automatic content analysis and the subjectivity of video summarisation. The proposed system allows a user to easily create a summarised movie clip composed of the most informative portions of the raw video. The rest of the paper is organised as follows. In Section 2 a description of the proposed home video summarisation system is presented. Section 3 gives a description of sub-shot segmentation, including an experimental analysis of sub-shot segmentation in our framework. Section 4 is devoted to a short description of the summarisation engine. The proposed user-interface design approach is then presented in Section 5. Finally, a conclusion is given in Section 6.

2 Proposed Home Movie Summarisation System

Our home movie summarisation system comprises a number of automatic visual-content analysis techniques as well as a user interaction step as shown in Figure 1. A raw video is fed into the sub-shot segmentation module, which computes the global camera motion parameters of the video and in turn decomposes it into 4 different sub-shot types called pan, tilt, zoom and static, as described in Section 3. Each identified sub-shot is then further processed using the content representation method described in [8]. The resulting sub-shot footprints are input to the summarisation engine, which then performs analysis to identify the most dynamic content as well as the redundant information present in the raw video footage. User interaction functionalities are supported to drive the summarisation process whereby a user will be able to create a particular summary of desired target length whilst being able to interactively view and refine the summarisation results (see Section 5 for details of the user interaction features supported by the system). If the user is happy, she can then confirm and request the final summary be created and saved to disk.

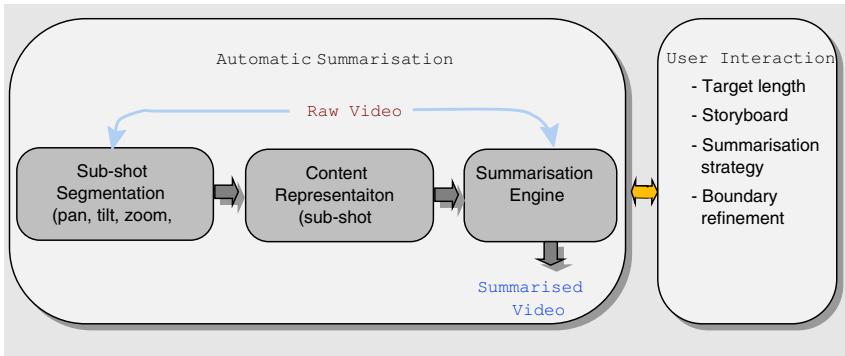


Fig. 1. Proposed home movie summarisation system

In this paper, we focus primarily on design concepts of the graphical user interface as other aspects of the system are reported elsewhere [8][14]. Additionally, we present a description of the experimental analysis carried out on sub-shot segmentation, emphasizing the need for identifying an efficient and robust sub-shot segmentation method for home movie summarisation.

3 Sub-shot Segmentation

Employing an effective sub-shot segmentation method is crucial to the success of home video segmentation, which in turn can lead to a significant reduction of subsequent user-interaction required for creating a visually appealing summary for real users. Sub-shot segmentation of home movie footage based on the use of camera motion estimation is, however, a highly computationally expensive task [8][15]. Thus, identifying an efficient and robust sub-shot segmentation technique is particularly important for the proposed summarisation framework. Although other techniques of much lower computational complexity appear to exist in the literature, we believe that detecting sub-shots in line with the change in dominant camera motion enables us to uncover the structure of raw home movie content more effectively.

The flow diagram of the sub-shot segmentation approach employed in our framework is shown in Figure 2. In the pre-processing stage, the raw video (V_r) is first decoded following which each frame of the video is converted to grayscale and resized. Then, the optical flow field is computed for each pair of consecutive frames in the camera motion estimation stage. Fitting those motion vector fields to a 2-D affine model and combining with the RANSAC algorithm, the best transformation between each pair of frames is computed. By comparing the values of each model parameter with a suitably determined threshold, classification of the global motion is carried out for each frame of the video. Finally, a filtering step is applied to all classified frames to determine the type of sub-shots present in the video.

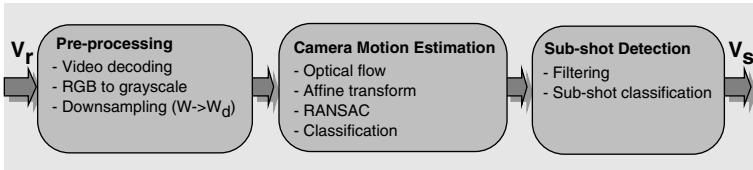


Fig. 2. Flow diagram of sub-shot segmentation: V_r and V_s represent the raw and sub-shot video respectively

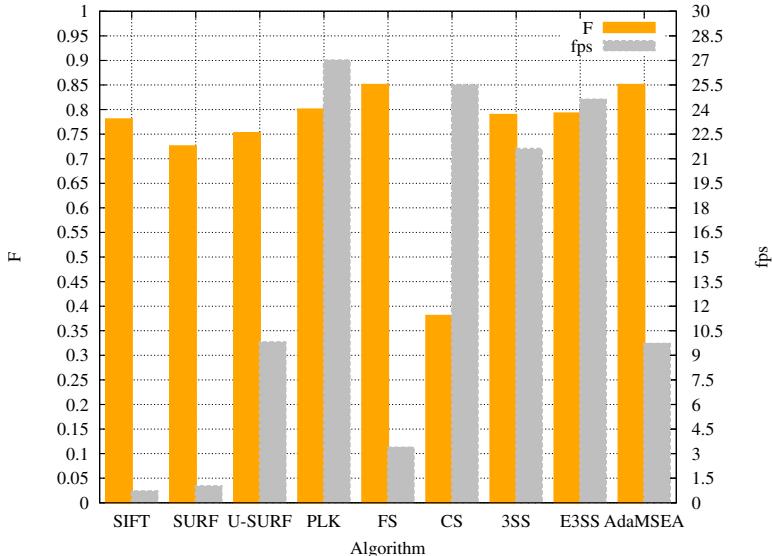


Fig. 3. Performance comparison of sub-shot segmentation based on the use of different optical flow field computation methods

Based on our extensive experiments carried out on home movie summarisation, it was evident that computing the optical flow field corresponds to the most time consuming task in sub-shot segmentation. For example, using SIFT features for sub-shot segmentation results in over 90% of the total computational time of summarisation [8]. To this end, we conducted experiments to identify an efficient and robust sub-shot segmentation technique based on the use of different feature-based and block-matching optical flow field computation methods. Feature-based methods include SIFT [16], SURF [17] and Pyramidal Lucas-Kanade (PLK) [18] while block-matching algorithms include full search (FS), cross search (CS) [19], three step search (3SS) [20], efficient 3SS (E3SS) [21] and adaptive multilevel successive elimination algorithm (AdaMSEA) [22]. A test data set consisting of 28 typical home movies (191 sub-shots in total) with varying camera motions, events and durations was used to test the performance of sub-shot segmentation in this experimental study.

Figure 3 shows a comparison of the different algorithms in terms of accuracy against efficiency using the F-measure and frames-per-second (fps)

evaluation criteria. It can be seen that the full-search block matching algorithm and AdaMSEA (which falls into the category of efficient full-search block algorithms) perform best, followed by PLK, E3SS, 3SS, SIFT, U-SURF, SURF and CS in terms of accuracy. There is a 0.05 difference in “F” value between the two best performing algorithms, i.e. full-search (or AdaMSEA) and PLK. However comparing the efficiency figures of AdaMSEA makes it clear that, considering the accuracy/efficiency compromise home movie summarisation applications require, the PLK algorithm is the preferred optical flow computation choice for sub-shot detection in home movie content. A detailed performance comparison of the algorithms is given in [14].

4 Summarisation Engine

The summarisation engine performs a range of automatic analysis to facilitate the creation of a summarised video. Relying on the content representation method built on the principle of sub-shot footprints [8], the main function of the summarisation engine is to automatically identify which portions of the raw video to be included in the summary, given the user inputs such as target length, summarisation strategy, coverage, etc. It also provides additional information to the user during user interaction stages, such as manually adjusting the boundaries of video segments, changing the target length, and selecting the summarisation strategy.

In order to rank a set of sub-shots based on the significance of content and select the most relevant sub-shots for summarisation, the summarisation engine analyses the coverage and intersection of sub-shot footprints [8]. Based on this analysis, our summarisation framework supports three different summarisation

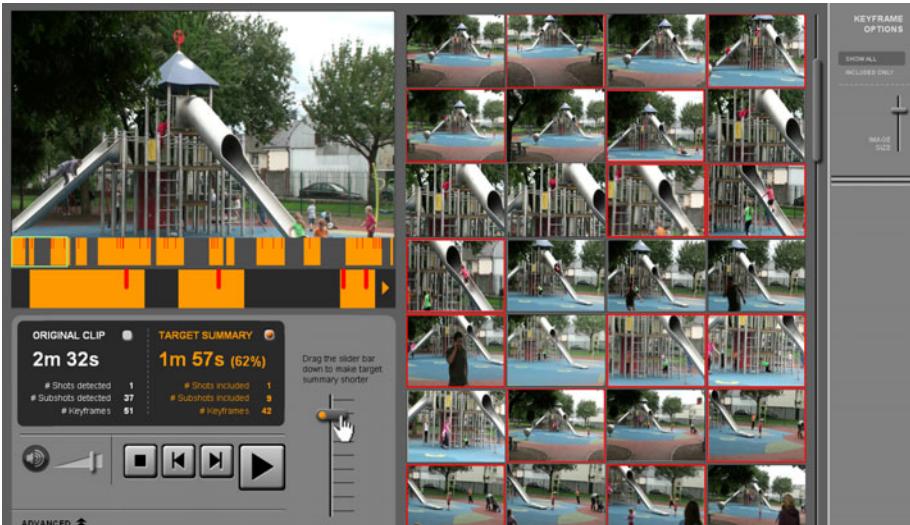


Fig. 4. Screen-shot showing initial home movie summarisation results

strategies, thereby providing users the flexibility to select the best automatically edited version of a video which can be subsequently refined to tailor for their own needs with minimum interaction. Definitions of the three summarisation strategies provided by the system through the use of an advanced panel (see Figure 5) are as follows. Given target length, T:

- *Strategy 1 (Prominent)*: Arrange sub-shots iteratively so that those corresponding to high coverage and low intersection appear in the top positions of the rank list until their aggregate length ie equal to T.
- *Strategy 2 (Coverage)*: Arrange sub-shots iteratively so that those corresponding to high coverage and low intersection appear in the top positions of the rank list until a pre-defined level of coverage is reached. Then, extract the most dynamic segment from each of those sub-shots so that their aggregate length is equal to T.
- *Strategy 3 (All)*: Extract the most dynamic segment from each of the full set of sub-shots so that their aggregate length is equal to T. This option is set as the default setting, ensuring that every sub-shot is preserved to some degree in the final summary.

5 Interaction Design

Existing video editing tools, such as iMovie and Adobe Premiere, allow users to perform video editing as a fully manual process. They also inevitably feature complex user-interfaces with a number of layers of timelines to be overlayed, expecting the user manipulates this by adding more segments, adjusting the temporal sequences between the segments, etc. The strength of the proposed video summarisation system is that it simplifies the manual manipulation process by automatically preparing a near optimal summary template, which the user can further refine to tailor for his/her own needs with simple interaction.

The scenario and the front-end user-interaction we have designed is to support a user to easily view the contents, automatically summarise into a compact video clip, and then manually adjust if wished. In this way, automatic summarisation could serve as a pre-edit processing that takes care of most of the otherwise time-consuming, repetitive and labour-intensive editing tasks.

5.1 Initial Summarisation and Browsing Scheme

Figure 4 shows a screen-shot where a home video of a user and her children playing in a playground is loaded into the system, and initially processed and presented to the user on her laptop screen.

On the top-left of the screen is the video playback panel where the user can play the original or summarised video. Below it is a timeline that shows the segments of the video that are included (in orange) and omitted (in dark grey) in the summary. Thin vertical bars marked on the top edge of the timeline (in red) indicate the positions in the original video from where the representative keyframes have been extracted. The timeline is double-layered where the top half



Fig. 5. Advanced user control of the summarisation

shows the full duration of the raw video providing an overview of the summary while the bottom half shows a zoomed-in portion taken from the top half covering the green rectangular area (in this scenario the first 20 seconds of the original video). Presented on the right half of the screen is a “storyboard” of static keyframes. The keyframes marked with red borderlines correspond to the thin red vertical bars on the timeline. The user has an option to view only those selected keyframes (highlighted in red) by selecting the “Included Only” option on the Keyframe Options panel on the right. Also, the size of each keyframe can be changed by dragging up and down the “Image Size” vertical slider bar below.

The current status of summarisation is indicated in the dark text box just below the timeline. In Figure 4, this text box indicates that a “2 minutes and 32 seconds” long raw video has been summarised into a shorter duration of “1 minute and 57 seconds” corresponding to 62% of the original size. At this stage the user can click on the Play button (the larger of the 4 buttons provided below the text box) to view the resultant summary on the player screen that will only play the orange-highlighted parts on the timeline. The other 3 buttons allow the current playback point to jump to the beginning of the next and previous selected segment or stop. Button actions are applied to the summarised video as the yellow radio button beside the label “Target summary” in the text box is selected. When the radio button for the original clip is selected, the buttons that jump to next/previous segment switch to Fast Forward/Backward buttons, in order to allow quicker navigation of the original video clip.

While the above is the initial summarisation outcome automatically generated by the system, the user can now customise the summary in different ways. The vertical slider bar provided just beside the 4 buttons can be dragged up and down in order to change the target length of the summary. For example, as the user

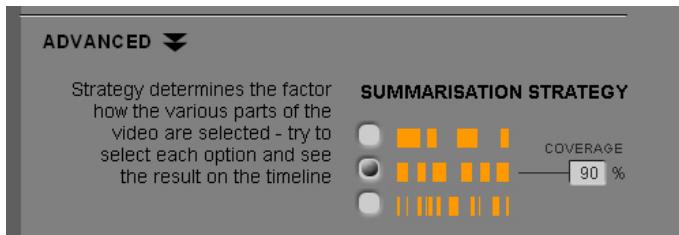


Fig. 6. Selecting a summarisation strategy using advanced panel

drags the knob down, she will immediately notice the orange-highlighted portions on the timeline becoming thinner and thinner, the number of keyframes with red borders in the storyboard decreasing, and the target duration displayed in the text box becoming smaller. Given a target length, the summarisation engine then performs automatic analysis to determine the most informative portions of the original video to be included in the summary. In this way, our system facilitates users creating video summaries through a simple user intervention process while making best use of the automatic content analysis technologies.

5.2 Advanced Summarisation

A more fine-grained user control of the summarisation process is possible if the user wishes. The “Advanced” button provided at the bottom-left of the screen can be clicked to bring up a small panel where the user can specify different summarisation strategies. Figure 6 shows a screenshot where the Advanced panel has been slid up, and the user dragged down the vertical slider bar in order to reduce the summary duration to only 32 seconds.

A detailed view of the advanced panel is shown in Figure 6, depicting the options available for “summarisation strategy”. The 3 options represent (1) include prominent sub-shots, (2) select sub-shots based on coverage, and (3) include as many of all sub-shots. Option 2 is accompanied by an additional setting for the *coverage* where the default value is set to 90%. If the user selects either option 1 or 3, the Coverage setting will be disabled (see Figure 5). Even if the user does not fully understand the technical implications of these settings, she can quickly experiment with these options by observing the immediate changes on the timeline and the storyboard. In general, the user can notice that option 1 tends to generate a summary with a small number of large chunks of the video; a medium number of medium sized chunks for option 2; and for option 3 a large number of small chunks. Any of these options can be used in conjunction with the adjacent vertical slider bar that lets the user set the duration of the summary.

5.3 Summary Customisation: Manual Refinement

If the user is not happy with the above results, she can further customise the summaries through manual refinement. She can simply move the mouse cursor over the timeline and drag the borders between orange (selected portion) and



Fig. 7. Manually refining the boundaries of a segment

dark gray (omitted portion) to manually adjust the boundaries of the video segments. Figure 7 shows an enlarged timeline where the user initially dragged the green rectangular frame on the top half of the timeline to about 3/4 into the video with the bottom half showing a zoomed-in portion of that area. The user is currently extending the orange block by dragging its border to the right, which is indicated by the lighter orange colour area as the portion being added to the summary. As the user adjusts the segment's boundary in this way, she can see the relevant changes being displayed in the text box and the storyboard. In Figure 5, the storyboard highlights 21 keyframes (only 6 of them shown in the top part of the scrollable storyboard) as a result of this adjustment as opposed to the initially highlighted 42 keyframes (and 14 of them shown) in Figure 4.

6 Conclusion

We presented a novel home-movie summarisation framework that combines automatic content analysis with an intuitive user-interface design approach, exploiting the synergy between computer power and human abilities to guarantee the best summarisation results in a simple and efficient manner. We demonstrated the challenges associated with automatic video segmentation through an experimental analysis of sub-shot detection in home movie footage. Using the combined approach, we propose that the issues arising due to the subjectivity of video summarisation and automatic content analysis can be suitably addressed. At present, the proposed user-interface exists at the design stage. By employing an intuitive user-interface design approach that functions in combination with efficient content analysis modules in the back-end, we believe that our approach meets the requirements of real home-movie users. We intend to carry out a number of evaluation tests to assess the effectiveness of our approach based upon the feedback from real users in the future.

Acknowledgment

This research is supported and funded by *BT Group plc*, *IRCSET* and the *Science Foundation Ireland* under grant 07/CE/I1147.

References

1. Lienhart, R.: Abstracting Home Video Automatically. In: ACM Multimedia, Orlando, FL, USA, pp. 37–40 (1999)

2. Kender, J.R., Yeo, B.-L.: On the Structure and Analysis of Home Videos. In: Proc. of ACCV, Taipei (January 2000)
3. Gatica-Perez, D., Loui, A., Sun, M.-T.: Finding Structure in Home Video by Probabilistic Hierarchical Clustering. IEEE Tran. on Circuits and Systems for Video Tech. 13(6), 539–548 (2003)
4. Huang, S.-H., Wu, Q.-J.: Intelligent home video management system. In: Intl. Conf. on Information Technology: Research and Education, pp. 176–180 (2005)
5. Mei, T., Hua, X.-S., Zhou, H.-Q., Li, S.: Modeling and Mining of Users' Capture Intention for Home Videos. IEEE Tran. on Multimedia 9, 66–77 (2007)
6. Takeuchi, Y., Sugimoto, M.: User-Adaptive Home Video Summarization using Personal Photo Libraries. In: Proc. of CIVR, pp. 472–479 (2007)
7. Wang, P.P., Wang, T., et al.: Information Theoretic Content Selection for Automated Home Video Editing. In: Proc. of ICIP, Texas, USA, pp. 537–540 (2007)
8. Cooray, S.H., Bredin, H., Xu, L.-Q., O'Connor, N.E.: An Interactive and Multi-level Framework for Summarising User Generated Videos. In: ACM MM, Beijing, China, pp. 685–688 (2009)
9. Peng, W.-T., Huang, W.-J., et al.: A User Experience Model for Home Video Summarization. In: Proc. of MMM, Chongqing, China, pp. 484–495 (2009)
10. Grgensohn, A., Boreczky, J., et al.: A Semi-automatic Approach to Home Video Editing. In: Proc. of ACM Symp. on User Interface Software and Technology, San Diego, CA, USA, pp. 81–89 (November 2000)
11. Campanella, M., Weda, J., Barbieri, M.: Edit while watching: home video editing made easy. In: Proc. of SPIE, vol. 6506 (2007)
12. Wu, P., Obrador, P.: Personal Video Manager: Managing and Mining Home Video Collections. In: Proc. of SPIE, Bellingham, vol. 5960, pp. 775–785 (2005)
13. Salton, G., Singhal, A., et al.: Automatic Text Structuring and Summarization. Information Processing and Management 22(2), 193–207 (1997)
14. Cooray, S.H., O'Connor, N.E.: Identifying an Efficient and Robust Sub-shot Segmentation Method for Home Movie Summarisation. Accepted for publication in 10th IEEE Intl. Conf. on Intelligent Systems Design and Applications (ISDA), Cairo, Egypt, November 29 - December 1 (2010)
15. Tang, L.-X., Meo, T., Hua, X.-S.: Near-Lossless Video Summarisation. In: ACM MM, Beijing, China, pp. 1049–1052 (2009)
16. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Intl. Journal of Computer Vision, 91–110 (2004)
17. Bay, H., Ess, A., et al.: SURF: Speeded Up Robust Features. In: Computer Vision and Image Understanding (CVIU), vol. 99(3), pp. 346–359 (2008)
18. Bouguet, J.-Y.: Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm, Part of OpenCV library
19. Ghanbari, M.: The Cross-Search Algorithm for Motion Estimation. IEEE Tran. on Communications 38(7), 950–953 (1990)
20. Koga, T., Linuma, K.: Motion Compensated Interframe Coding for Video Conferencing. In: Proc. Nat. Telecommunication Conf., pp. G5.3.1–G5.3.5 (1981)
21. Jing, X., Chau, L.-P.: An Efficient Three-Step Search Algorithm for Block Motion Estimation. IEEE Tran. on Multimedia 6(3), 435–438 (2004)
22. Liu, S.-W., Wei, S.-D., Lai, S.-H.: Fast Optimal Motion Estimation Based on Gradient-Based Adaptive Multilevel Successive Elimination. IEEE Tran. on Circuits and Systems for Video Technology 18(2), 263–267 (2008)

Image Super-Resolution by Vectorizing Edges

Chia-Jung Hung, Chun-Kai Huang, and Bing-Yu Chen

National Taiwan University

{fffantasy1999, chinkyell}@cmlab.csie.ntu.edu.tw,
robin@ntu.edu.tw

Abstract. As the resolution of output device increases, the demand of high resolution contents has become more eagerly. Therefore, the image super-resolution algorithms become more important. In digital image, the edges in the image are related to human perception heavily. Because of this, most recent research topics tend to enhance the image edges to achieve better visual quality. In this paper, we propose an edge-preserving image super-resolution algorithm by vectorizing the image edges. We first parameterize the image edges to fit the edges' shapes, and then use these data as the constraint for image super-resolution. However, the color nearby the image edges is usually a combination of two different regions. The matting technique is utilized to solve this problem. Finally, we do the image super-resolution based on the edge shape, position, and nearby color information to compute a digital image with sharp edges.

Keywords: super-resolution, vectorization, matting, edge detection, Bézier curve, mean-value coordinate, interpolation.

1 Introduction

Image super-resolution is a task that scales a digital image up. The ability of scaling up a digital image is very important for many aspects. For example, there are more and more high-definition display devices but not all contents produced in such high resolution, so we have to scale these contents to fill up the whole display.

Super-resolution is a very ill-posed problem due to its nature. If we want to get a high resolution image with 2x large width and 2x large height, only 1/4 pixels of the target image can be obtained from the original image perfectly. Other 3/4 pixels cannot be determined uniquely. To regularize this problem, we have to make some assumptions.

The most commonly used assumption is that the image is locally smooth. According to this assumption, many interpolation based methods have been proposed. Three well-known interpolation methods are nearest neighbor, bilinear interpolation, and bicubic interpolation. However, these three methods would produce some unwanted artifacts as shown in Fig. 1 (b), such as the result image may be blurry and textureless and the edges in the result image may be jaggy or blocky. Because of these problems, many algorithms have been proposed to solve some parts of them.

In this paper, we focus on the image super-resolution while preserving the image edges, because image edges are strongly related to human perception of image

quality. We consider that blocky artifact would decrease the image quality most seriously. Inspired by image vectorization techniques, we noticed that if we can represent the image edges by some parameterization methods. We can reproduce them at any resolution with the preserved edges. Hence, in this paper, we try to extract the image edges and use a parametric representation to capture them. We can get an enlarged image with these edge data while preserving the edges.

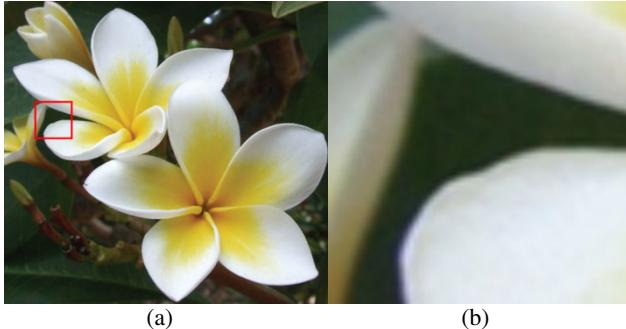


Fig. 1. The result of bicubic interpolation, where (b) is the enlarged red square of (a)

2 Related Work

Super-resolution has been an interesting topic for a long time, so there are many different algorithms have been proposed. Since it is an under constraint problem, two typical approaches are usually used to overcome this problem, which are adding data and adding constraint.

To add more data, multiple image super-resolution raised. They use multiple low resolution images of the same scene with sub-pixel displacement as the input to compute a high resolution one. Single image super-resolution includes a wide range of work. As summarized in [14].

In recent researches, [16] proposed a method that builds an over-complete dictionary of low resolution image patches from a large image set, and uses a sparse representation of the image with the dictionary to do the super-resolution. [5] proposed a new image prior using image gradients and used these gradients learned from a bunch of natural images to estimate a high resolution image from a low resolution counterpart. [12] is similar to [5], and based on the statistics about the prior it can produce a natural high resolution image.

[8] and [11] are very similar to the anisotropic diffusion. They scale the image up via an interpolation base method, and try to sharpen the edges. Anisotropic diffusion directly employees the well known image sharpen algorithm “anisotropic diffusion”, while [8] and [11] deblur the scaled blurry image.

Besides, [13] proposed a tensor voting mechanism to do the super-resolution, and [3] proposed a soft edge prior to do it while preserving the edges and keeping the smoothness of each edge. Generally speaking, algorithms that take image edges into consideration can produce a more satisfactory image.

3 System Overview

Our algorithm is derived from bilinear interpolation and image vectorization. Basically, Image edges stand for a large color difference, so to sample the colors from different sides of an edge for interpolation would produce jaggy edge. To do the interpolation without edge crossing, we first vectorize the image edges, and analyze the color compositions to get a more compact representation of the image edges.

Fig. 2 shows the system flowchart. There are several main components in our system, which are:

- Edge Detection and Edge Extraction;
- Matting based Image Color Analysis;
- Sub-pixel Refinement and Edge Shape Fitting;
- Represent the image using a Polygonal Representation;
- Edge-Preserving Super-resolution.

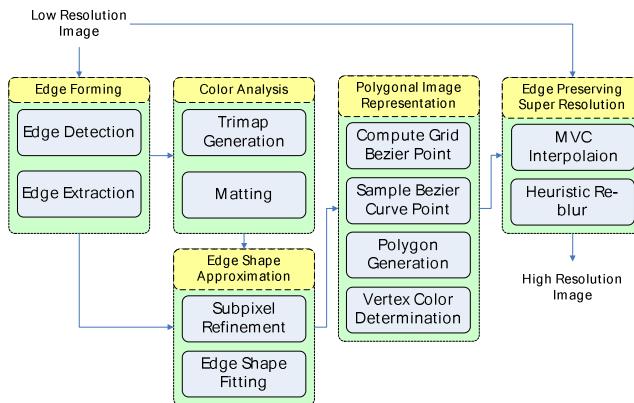


Fig. 2. System flowchart

Since we want to vectorize the image edges, the well known Canny edge detector [1] is used to compute the edge map. The detected edge pixels are linked to form the edges. After extracting the edges from the edge map, we analyze the color information nearby the edges by using a matting algorithm. Then, we use these color information to improve the position of the edge pixels and record these color data as a component of the associated edges. As long as sub-pixel refinement is done, we can vectorize the edge with piecewise smooth cubic Bézier curves. Bézier curve is a parametric curve, we can scale it to any resolution we want without loss its smoothness and any outline deformation.

Finally, to do the interpolation with the edges as the color sampling constrain, we employee the mean value coordinates (MVC) [6] to do the interpolation. MVC is a coordinate with only the function values defined on the polygon vertices. Hence, we make a polygonal representation of the original image with its pixel grids and Bézier curve samples as the vertices. Then, we do a heuristic Gaussian reblurring on the MVC interpolated image.

4 Edge Forming

4.1 Edge Detection

We use the Canny edge detector [1] to find the edge pixels and employ the MATLAB version of the Canny edge detector, so that we can thin the edge to 1-pixel width successfully. Canny edge detector can only accept a single channel image to compute the edge map. We first convert the image from RGB color domain to YUV color domain, and only use the Y-channel image to compute the edge map. Because the edges detected in the Y-channel image are more intuitive for human.

4.2 Edge Extraction

After detecting the edge pixels, we link each pixel with its 8-way neighborhood. The edge map treats as a graph. For each pixel, we record its neighboring amount N , which indicates the degree of each pixel after edge extraction. First, we search all pixels with only one neighbor ($N = 1$) as the roots, then traverse the map in a DFS manner. As a pixel has been connected, we decrease its neighbor amount N to reflect how many times the pixel linked.

The graph is traversed from each root until we meet another pixel that has only one neighbor or a pixel with $N = 0$, and this path forms an edge. Because, most edge pixels can have only two neighbors, this process can traverse most edges without any problem. However, if we encounter a pixel has more than two neighbors, we choose the one with smaller spatial distance and color distance as the next pixel. After we traversed all edges from 1-neighbor root, we search all 2-neighbor pixels as the roots again, while the procedure is the same.

5 Edge Color Analysis

After edge extraction, we want to interpolate the pixel color without crossing the edges. We need to do the interpolation on a target pixel that is nearby some Bézier curves, we will use the color samples from the Bézier curves and those from the grid points at the same side of the edge to compute its color and preserve the edge at the same time.

5.1 Trimap Generation

To utilize the image matting technology, a trimap is necessary. In our system, we generate it automatically by assigning one side of the edge as the foreground region and another side as the background region. However, the edge can reside anywhere in the image, and it may only separate a small area of the image into two regions. Therefore, to generate a trimap associated to an edge, we have to crop a patch of the image nearby the edge first, and then solve the colors by image matting.

5.2 Matting

After generating the trimap of each edge, we are ready to solve the image matting problem. Though the generated trimap is not so perfect, the closed-form matting [9] can generate quiet adequate solution for most cases. As shown in Fig. 3, (a) is the

cropped images for each edge extracted by our algorithm; (b) is the automatically generated trimaps (the white, black, and gray colors indicate the foreground pixels, background pixels, and unknown region, respectively); (c) is the alpha map solved by closed-form matting [9]; (d) and (e) are the solved background and foreground images, respectively.

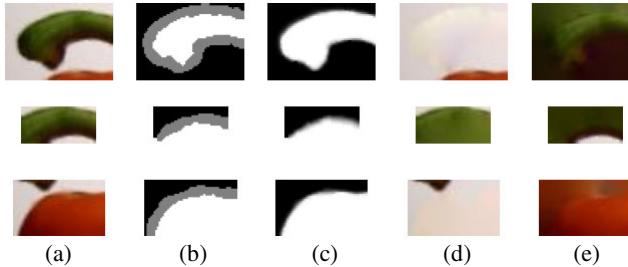


Fig. 3. Result of trimap generation and matting algorithm

6 Edge Shape Approximation

6.1 Sub-pixel Refinement

An ideal edge should reside in between the gradient local maximum and local minimum. The Canny edge detector only has pixel-level precision. Furthermore, an edge in a color image should relate to all color channels. If we detect each color channel separately, how to merge them will be a problem due to the inconsistency. To overcome this problem, we use an alpha map generated by the matting algorithm. As [13] depict, the alpha values are adequate to do the edge pixel enhancement.

Rather than simply utilizing the sub-pixel refinement method in [13], we compute the sub-pixel position of an edge pixel by a method similar to Harris corner detector [7]. We think that the ideal position of an edge pixel should between the foreground and background that means an edge pixel should have $\alpha = 0.5$. To find such position, we slice the alpha value along the gradient of the point as an 1D function, and search a position of $\alpha = 0.5$ approximately. First, we approximate the 1D function using the Taylor expansion:

$$f(x) \approx f(0) + f'(0)x + \frac{f''(0)}{2}x^2.$$

Then, we can solve $f(x) = 0.5$ by the above formula and moving the edge pixel to x .

6.2 Edge Shape Fitting

After extracting the edges from the edge map and the sub-pixel refinement, we can fit the edge shape by a piecewise smooth cubic Bézier curve. For an edge with point P_0, P_1, \dots, P_n , we want to find a piecewise Bézier curve $Q(t, V)$ that fits P_0, P_1, \dots, P_n . Assume that the curve $Q(t, V)$ passes through P_0 and P_n , it could be defined as:

$$Q(t, V) = \sum_{k=0}^3 B_k(t) V_k ,$$

where $0 \leq t \leq 1$, $V_0 = P_0$, $V_3 = P_3$, $V = (V_1, V_2)$, and

$$B_k(t) = \binom{3}{k} t^k (1-t)^{3-k} .$$

Then, we can try to find the curve $Q(t, V)$ by minimizing

$$D(t, V) = \sum_{i=1}^{n-1} d_i = \sum_{i=1}^{n-1} [Q(t_i, V) - P_i]^T \cdot [Q(t_i, V) - P_i] ,$$

where $t = (t_1, \dots, t_{n-1})$.

We employ the algorithm in [2] to do the curve fitting. When the average fitting error of a curve exceeds a threshold set by the user (in our experiments, we set it to 0.5), we split the curve into two curves at a point with the largest fitting error. We do not force smoothness in the conjunction point of the edge split; because of keep it unsMOOTH can preserve its shape better.

7 Polygonal Image Representation

Because the edges' neighborhood may be overlapped and we need a global pixel value when we calculate the target image. To do the interpolation without edge crossing, we use MVC (Mean Value Coordinate) [6] to interpolate the pixel values by using the original image grids and the Bézier curve points as the MVC polygons' vertices.

7.1 Computing Bézier Grid Points

To get the Bézier grid points on the original image grids means that we want to find the following set:

$$S = \{t \mid Q(t, V).x \in Z \wedge Q(t, V).y \in Z\} .$$

That means we have to solving t by given $Q(t, V)$ and an integer n . Basically, it is a root-finding problem; however, since our equation is in the Bézier form, it can be solved by a more efficient method called Bezier clipping [10].

7.2 Sampling Bézier Curve Points

After all Bézier grid points of one Bézier curve are calculated, the set S is sorted for further usage. When we get all grid samples from an edge, we uniformly sample the points between two consequent grid points by simply interpolating the parameter between those points.

7.3 Polygonal Image Representation

Because there can be multiple polygons in an original image grid and the MVC needs polygon vertices in counter clockwise, we build an association list that uses the original image grids as the indices, and record which Bézier grid point belong to the image grid point's neighboring. According to this association list, we can form a point

list by traversing each grid point's list counter clockwise. Then, we build a polygon list from the point list.

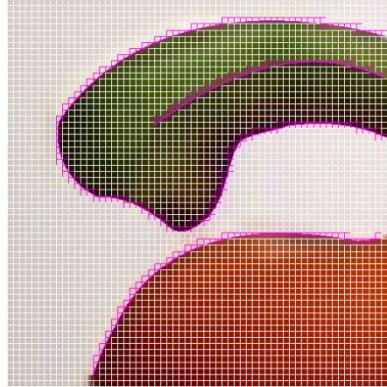


Fig. 4. An example of polygonal representation of an image

Note that the Bézier point in between the two Bézier grid points can reside outside of the current image grid, so before connecting them, some checks must be done. Because we use every point inside the list as a starting point of a polygon, we have to check the polygon before inserting it into the polygon list. Finally, we can get the polygon list of a grid.

7.4 Vertex Color Determination

Because vertex color can be affected by the edges, we have to use the edge position as a hint to determine the color of each vertex. Hence, we scale the edges first, and then we can compute the vertex color. For each edge, we have assigned each side of it with different regions in the trimap generating step. To determine the color of each vertex, we have to scale the edges to the target resolution, and determine where the foreground and background regions are in the target resolution.

However, we have sampled each Bézier curve into a sequence of points in the polygon generating step. Therefore, we scale each sample point, and then connect the consequent points with a single line. At last, we use an identical foreground and background assignment of the trimap generating step by similar rules. In the following section, we call the scaled foreground and background map as FBMap.

For all Bézier curve points, we assign its color as blend and record its associated Bézier curve, and determine its color until we do the interpolation. For a vertex of the original image grid, it can be covered by FBMap. Here we say “cover” means that the vertex resides in the unknown region of the trimap of the associated edge. If there is no FBMap covers it, we will use the original image pixel color as its color. There is only one FBMap covers it. If it belongs to the foreground or background region, we assign it the color of the foreground or the background. If it belongs to the blend region, we have to determine its color until we do the interpolation,. If there are more than one FBMap cover it, we can calculate the pixel color within each FBMap by fore mentioned method, and calculate an associated confidence value defined by [15]:

$$R_d(F, B) = \frac{\|C - (\alpha F + (1-\alpha)B)\|}{\|F - B\|},$$

where C is the original pixel color, α is the alpha value, and F and B are foreground and background colors, respectively. We choose the pixel color with the lowest confidence value as its color.

8 Edge Preserving Super Resolution

8.1 Mean Value Coordinate

While taking the edges as the constraint, we use the MVC to interpolate the pixel colors. After we determine the polygonal representation of the image, we can use the MVC to interpolate the pixel colors inside each polygon smoothly.

8.2 Image Interpolation Using MVC

For each scaled grid of the image, we first count how many polygons inside the grid. If there is only one polygon inside, we can directly apply the MVC with the vertices' pixel colors as the function values to do the interpolation. However, we use the bilinear interpolation to accelerate the process. If there is more than one polygon inside, we need to determine the target pixel belongs to which polygon and then use the polygon to do the MVC interpolation.

When we do the interpolation, there are still some vertices' colors have not been determined, so as we find which polygon the target pixel belongs to, we have to examine the color of each vertex. If the vertex color has been assigned blend, then we first determine the polygon belongs to which side of the associated Bézier curve. If the polygon belongs to the foreground, then each vertex with the blend color should be assigned the foreground color of the edge and vice versa.

8.3 Image Reblurring

When we sample the color that affected by the edge, we use the foreground and background colors directly. This procedure makes the gradient along the edge of our system becomes a step function that contains only the pure foreground and background colors as shown in Fig. 5 (a). For a nature image, the gradient along the edge should be a smooth function. As [4] depict, this phenomena will make the image unnatural, so reblurring is needed for a more natural image.

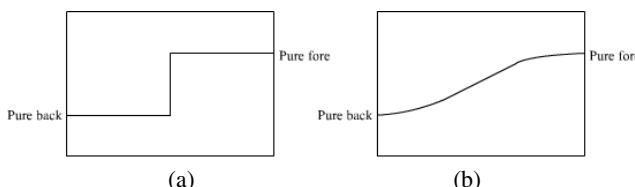


Fig. 5. Function along edge gradient

9 Result

In this section, we show some results of our method. In each of our experiment, we scale the original image to 8x size. Fig. 6 shows the results. Table 1 lists the performance of each step of our system. We tested our system on a desktop PC with an Intel Core2Quad 2.4GHz CPU with 3.0GB RAM without any optimization. The performance depends on the edge extraction and the input image size.

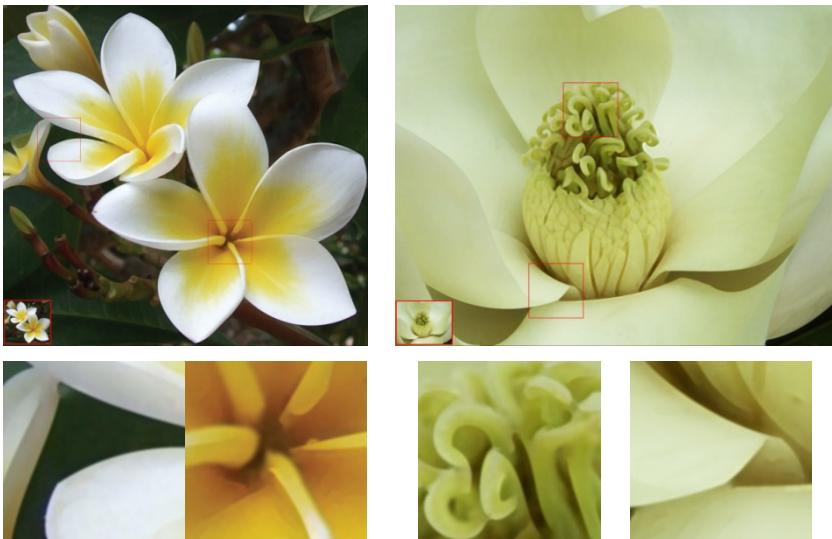


Fig. 6. Result images

Table 1. Performance

	Case1	Case2
Width (px)	511	535
Height (px)	397	500
Edge Number	325	435
Edge Detection (sec.)	0.75	0.83
Edge Generation (sec.)	0.01	0.03
Trimap Generation (sec.)	32.08	73.81
Mat Solving (sec.)	16.05	20.97
Sub-pixel Refinement (sec.)	0.14	0.13
Edge Shape Fitting (sec.)	10.06	22.16
Béizer Curve Sampling (sec.)	1.00	1.56
FBMap Generation (sec.)	12.14	33.77
Vertex Color Determination (sec.)	38.05	62.36
Edge Preserving Interpolation (sec.)	278.8	502.2
Total Running Time (sec.)	389	717

10 Conclusion and Future Work

In this paper, we proposed a new method to do the edge-preserving image super-resolution. We represent the image edges with an explicit parametric form, and use the image matting as a tool for color analysis. Our system can produce acceptable result even in a very large scale factor. However, since our algorithm is based on the Canny edge detector, there may be some failure cases due to it. To further enhance it is one of our future work. Besides the image quality, the performance is also a typical issue. There are some components may be accelerated by using SIMD instructions or GPGPU technologies.

Acknowledgement

This paper was partially supported by the National Science Council of Taiwan under 98-2622-E-002-001-CC2 and also by the Excellent Research Projects of the National Taiwan University under NTU98R0062-04.

References

1. Canny, J.: A Computational Approach To Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8(6), 679–698 (1986)
2. Chang, H., Yan, H.: Vectorization of Hand-Drawn Image using Piecewise Cubic Bézier Curves Fitting. *Pattern Recognition* 31(11), 1747–1755 (1998)
3. Dai, S., Han, M., Xu, W., Wu, Y., Gong, Y.: Soft Edge Smoothness Prior for Alpha Channel Super Resolution. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2007)
4. Elder, J.H.: Are Edges Incomplete? *International Journal of Computer Vision* 34(2-3), 97–122 (1999)
5. Fattal, R.: Image Upsampling via Imposed Edges Statistic. *ACM Transactions on Graphics* 26(3) Article no. 95 (2007)
6. Farbman, Z., Hoffer, G., Lipman, Y., Cohen-Or, D., Lischinski, D.: Coordinates for Instant Image Cloning. *ACM Transactions on Graphics* 28(3) Article no. 67 (2009)
7. Harris, C., Stephens, M.J.: A Combined Corner and Edge Detector. In: *Alvey Vision Conference*, pp. 147–152 (1988)
8. Joshi, N., Szeliski, R., Kriegman, D.J.: PSF Estimation using Sharp Edge Prediction. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2008)
9. Levin, A., Lischinski, D., Weiss, Y.: A Closed Form Solution to Natural Image Matting. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(2), 228–242 (2008)
10. Sederberg, T.W., Nishita, T.: Curve Intersection using Bézier Clipping. *Computer-Aided Design* 22(9), 538–549 (1990)
11. Shan, Q., Li, Z., Jia, J., Tang, C.K.: Fast Image/Video Upsampling. *ACM Transactions on Graphics* 27(5) Article no. 153 (2008)
12. Sun, J., Sun, J., Xu, Z., Shum, H.Y.: Image Super-Resolution using Gradient Profile Prior. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2008)

13. Tai, Y.W., Tong, W.S., Tang, C.K.: Perceptually-Inspired and Edge-Directed Color Image Super-Resolution. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 1948–1955 (2006)
14. van Quwarkerk, J.D.: Image Super-Resolution Survey. *Image and Vision Computing* 24(10), 1039–1052 (2006)
15. Wang, J., Cohen, M.F.: Optimized Color Sampling for Robust Matting. In: IEEE Conference on Computer Vision and Pattern Recognition (2007)
16. Yang, J., Wright, J., Ma, Y., Huang, T.: Image Super-Resolution as Sparse Representation of Raw Image Patches. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)

Vehicle Counting without Background Modeling

Cheng-Chang Lien¹, Ya-Ting Tsai¹, Ming-Hsiu Tsai¹, and Lih-Guong Jang²

¹ Department of Computer Science and Information Engineering,
Chung Hua University, Hsinchu, Taiwan, R.O.C.

Tel.: +886-3-5186404
cclien@chu.edu.tw

² Industrial Technology Research Institute, ISTC, Taiwan, ROC

Abstract. In general, the vision-based methods may face the problems of serious illumination variation, shadows, or swaying trees. Here, we propose a novel vehicle detection method without background modeling to overcome the aforementioned problems. First, a modified block-based frame differential method is established to quickly detect the moving targets without the influences of rapid illumination variations. Second, the precise targets' regions are extracted with the dual foregrounds fusion method. Third, a texture-based object segmentation method is proposed to segment each vehicle from the merged foreground image blob and remove the shadows. Fourth, a false foreground filtering method is developed based on the concept of motion entropy to remove the false object regions caused by the swaying trees or moving clouds. Finally, the texture-based target tracking method is proposed to track each detected target and then apply the virtual-loop detector to compute the traffic flow. Experimental results show that our proposed system can work with the computing rate above 20 fps and the average accuracy of vehicle counting can approach 86%.

Keywords: dual foregrounds fusion, texture-based target tracking.

1 Introduction

Recently, many vision-based researches addressed on the vehicle or human detection [1-5] was proposed. In general, the moving objects could be detected by three kinds of methods: motion-based [1], background modeling [2-4], and temporal difference [5] approaches. In the motion-based approaches, the optical flow method [1] utilizes the motion flow segmentation to separate the background and foreground regions. By applying the optical flow method [1], the moving objects can be extracted even in the presence of camera motion. However, the high computation complexity makes the real-time implementation difficult.

For the background modeling methods, the construction and updating of background models [2-4] often is time-consuming. For example, in [2,4], the Gaussian Mixture Model (GMM) is frequently adopted to model the intensity variation for each pixel over a time period and need high computing cost to calculate the GMM parameters. Furthermore, the foreground detection with background modeling method is extremely sensitive to the rapid illumination variation or the dynamic background changing. In [3], the Kalman filter is used to update the background model with less

computational complexity. But this method can't solve the problem of serious scene change which can make the system unable to update the background model accurately.

The temporal difference method doesn't need to establish the background model instead of subtracting the two adjacent frames and detecting the scene change introduced by the moving objects. The advantage of this method is less susceptible to the scene change, i.e., it has capability to detect the moving objects in dynamic environments. For example, in [5], the temporal difference method is resilient to the dynamic illumination variation, but the regions of the moving objects can't be extracted completely when the objects move slowly.

Here, we propose a novel vehicle detection method without background modeling to overcome the aforementioned problems. First, a modified block-based frame differential method is established to quickly detect the moving targets without the influences of rapid illumination changes. Second, the precise targets' regions are extracted by the dual foregrounds fusion method. Third, a texture-based object segmentation method is proposed to segment each vehicle from the merged foreground image blob and remove the shadows. Fourth, a false foreground filtering method is developed based on the concept of motion entropy to remove the false object regions caused by the swaying trees or moving clouds. Finally, the texture-based target tracking method is proposed to track each detected target and then apply the virtual-loop detector to compute the traffic flow.

The information of vehicle counting may be obtained from several kinds of vision-based sensor [6-10]. However, these methods are difficult to judge whether the vehicle appears in the virtual detector region or not under different lighting conditions. Furthermore, the lack of tracking information will make the vehicle counting inaccurate. Here, a two-line vehicle crossing scheme is applied to count the moving vehicles. The system block diagram is shown in Fig. 1. Our system consists of the following modules: vehicle extraction (foreground detection with dual foregrounds, foreground segmentation, and true object verification), vehicle tracking (Kalman filter tracking), and vehicle counting.

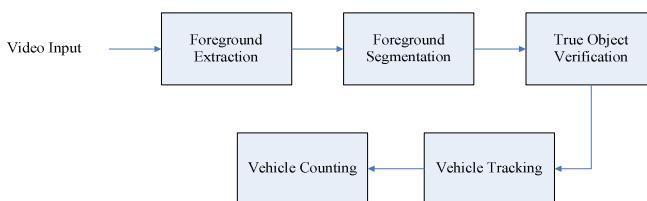


Fig. 1. The block diagram of the traffic flow analysis without background modeling

2 Vehicle Detection without Background Modeling

In general, vehicle detection can't be accurate on the light variation or cluster scenes. In this section, we propose a novel vehicle extraction method without the background modeling to segment the vehicles on the light variation or cluster scenes. Furthermore, to detect the vehicles efficiently and robustly, the motion-based false object filtering method is proposed.

2.1 Adaptive Block-Based Foreground Detection

Traditional pixel-based frame difference method can introduce many fragmented foreground regions because the spatial relationships among neighboring pixels are not considered. Hence, if we detect the moving objects using the frame difference method, both the local and global properties should be considered. Thus, we proposed the block-based foreground detection method to extract more complete foregrounds. First, each frame is divided into the non-overlapped blocks. The block size can be adjusted according to the object size. The foreground detection algorithm is described in the sequel.

1. We transform the RGB color space into the YC_bC_r color space and detect the moving object in Y channel, Cr channel, and Cb channel separately.
2. In each image block, if the number of detected foreground pixels exceeds a specified threshold, then we categorize this block into the foreground block.
3. By fusing the foreground regions detected from Y channel, Cr channel, and Cb channel with the voting rule, we can obtain a more complete object region.

The reason why we adopt the voting method is that Y, Cr, and Cb channels have different property in foreground detection. In Fig. 2-(a), the objects are detected with grayscale's variation in the Y channel; while in Fig. 2-(b)(c), the objects are detected with the color's variations in the Cr and Cb channels. The relationship between grayscale and color information are complementary.

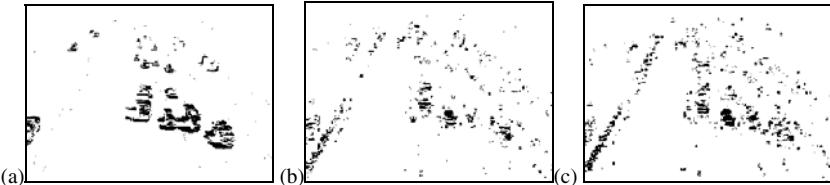


Fig. 2. The difference images for (a) Y channel, (b) Cr channel, and (c) Cb channel

If the region changes in the Y channel but doesn't change in the Cr and Cb channels, then it may be the object shadow. On the contrary, the region changes in the Cr and Cb channels but doesn't change in the Y channel, and then this region may be a noise region. Hence, in our system, the object detection with both the grayscale and color channels is established. The fusing rule is designed by the voting method. We adopted the rule-based method to determine whether the pixel/block belongs to foreground or not.

1. In each image block, if the difference of the Y channel exceed 1.5 times threshold of the Y channel, we classify this block into foreground directly.
2. Otherwise, the object is detected in the Y, Cr, and Cb channels with the rule: $(Y > T_Y) \&\& ((Cr > T_{Cr}) || (Cb > T_{Cb}))$. If the pixel/block changes in the grayscale and color channels obviously together, then we classify this pixel/block as foreground.

Through the careful observation, we found that the block-based method can extract the more complete foreground than the pixel-based method. When the block size becomes larger, the foreground becomes more complete and the computing time is also faster. However, the condition of wrong connection between different vehicles occurs more frequently.

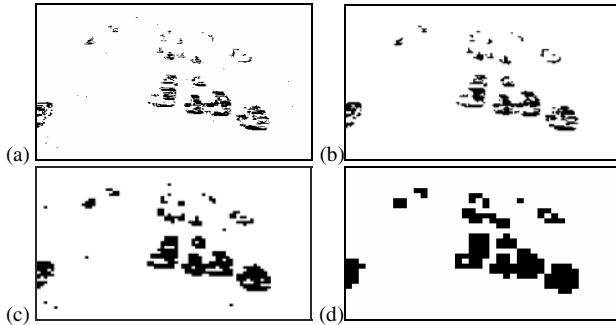


Fig. 3. Block-based foreground extraction. (a) Pixel-based. (b) Block-based (2×2). (c) Block-based (4×4). (d) Block-based (8×8).

Fig. 3 shows the comparison for the block-based and pixel-based method. It is obvious that the moving vehicles extracted with the block-based shown in Fig. 3-(b)(c) are more complete than the pixel-based method shown in Fig. 3-(a). In our experiment, we select the block size as 2×2 to fit the size of moving objects. So far, the block-based foregrounds belong to the short-term foregrounds that will be combined with the long-term foreground to generate a more precise and complete foreground, which will be described in next section.

2.2 Precise Object Region Extraction with the Dual Foregrounds

The traditional frame difference method often generate strong response in the boundary of moving object, but lower response occurs within the region of a moving object shown in Fig. 4-(b). When the object becomes larger, the incomplete object detection will become more serious. To tackle this problem we apply both the characteristic of short-term and long-term foregrounds to make the foreground more complete. The short-term foreground can define precise object boundary and the long-term foreground can extract a more complete object region shown in Fig. 4-(c) with motion history information. The long-term foreground is constructed by accumulating successive short-term foregrounds (In our experiment, the long-term foreground is constructed by accumulating 4 short-term foregrounds). By projecting the short-term foreground onto the long-term foreground, searching the precise object boundary based on the short-term foreground shown in Fig. 4-(f), and preserving all information about short-term and the long-term foregrounds between the boundaries of short-term foreground, we can extract the precise region of a moving vehicle shown in Fig. 4-(g).

2.3 Foreground Segmentation

In the outdoor environment, the serious shadow problem can introduce the inaccurate and false foreground detection shown as the object #4 in Fig.5-(b) and improper merging with neighboring objects shown as the object #2 in Fig. 5-(b). Based on the careful observation, the texture is not obvious in the shadow region illustrated as object #0 in Fig. 5-(b). The texture analysis is then used to eliminate the object shadows. Here, the texture analysis is performed by analyzing the gradient content obtained from the Sobel and Canny operations [11]. By projecting the edge information

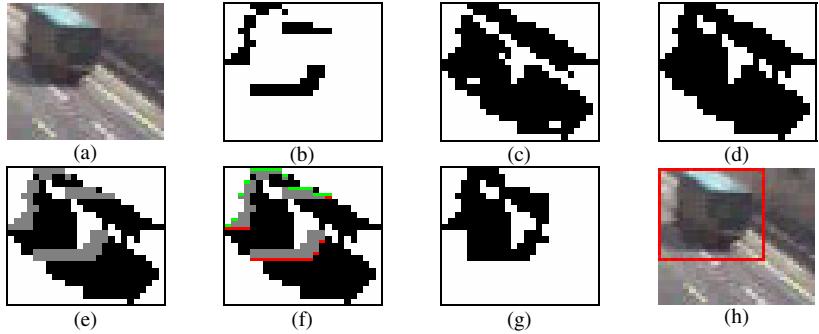


Fig. 4. Procedures for precise object region extraction with dual foregrounds fusion. (a) Original object. (b) Short-term foreground. (c) Long-term foreground. (d) Morphologic operation. (e) Integration of short-term (gray region) and long-term (black region) foregrounds. (f) Fusion with dual foregrounds (Red: upper bound; green: lower bound). (g) Object extraction by fusing the short-term and long-term foregrounds. (h) The bounding box for the extracted object in (g).

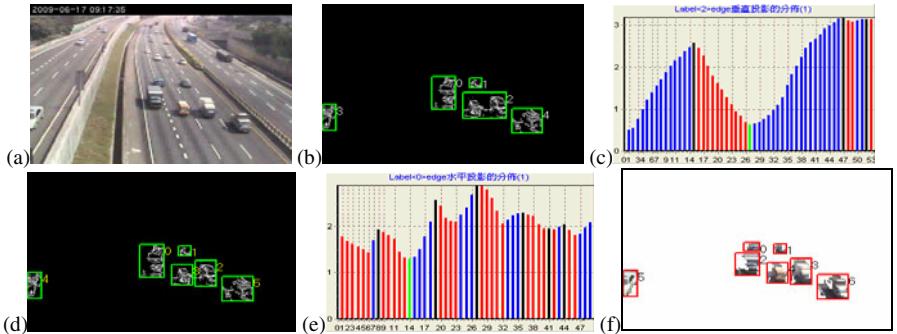


Fig. 5. The procedure of foreground segmentation. (a) Original image. (b) Canny edge detection and object labeling. (c) Edge projection to x-axis of object #2. (d) Object labeling after vertical segmentation. (e) Edge projection to y-axis of object #0. (f) Outcome of texture segmentation.

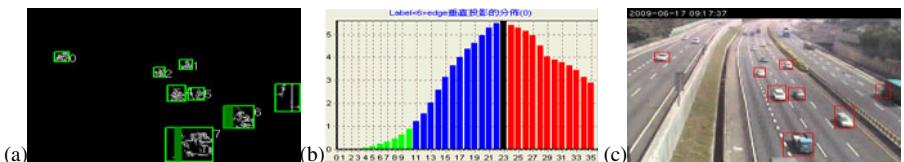


Fig. 6. Removal of the shadow region with the property of texture (a) The texture of foreground. (b) The texture density distribution for the 6th object region. (c) The result of shadow removal.

along horizontal and vertical axes, we can find the proper segmentation position (green bins) using the Ostu's method [12], which are shown in Figures 5-(c) and 5-(e). Once the merged vehicle region is partitioned, each precise vehicle region shown in Fig. 5-(f) is labeled via the labeling algorithm [13].

In general, the distribution of texture density of a moving object is higher than the distribution on its shadow region. Hence, we can utilize the texture densities to separate the moving object and shadow. By removing the boundary vertical regions with small texture density shown in Fig. 6-(b), the shadow region can be separated from the object region shown as the green regions in Fig. 6-(a).

2.5 True Object Verification

For a real moving object, the direction distribution of motion vectors should be consistent, i.e., the motion entropy in the object region will be low. Based on this observation, the value of motion entropy for each detected foreground region can be used to distinguish whether the detected object is a true object or not. First, we apply the three-step block matching method [14] to find the motion vectors for each foreground region and compute the orientation histogram of the motion vectors. Second, we compute the motion entropy for each detected foreground with Eq. (1).

$$E_m = -\sum_{i=1}^K p_m^i \log_2 p_m^i \quad (1)$$

where m is object index, i is the index of i -th bin in the orientation histogram, K is the number of total bins in the orientation histogram, and p_m^i is the probability of the i -th bin in the orientation histogram. The motion entropies in these false detected regions are very large and then these false detected regions can be removed with the motion entropy filtering process.

3 Vehicle Tracking

In this study, the texture feature that is not easily influenced by the illumination variation is utilized as the measurement in the target tracking algorithm.

3.1 Kalman Filter

In general, the detected vehicles can be tracked with the methods of Kalman filter or particle filters. With the efficiency consideration, we apply the Kalman filter [17] to track each detected vehicle on the roadway. Each detected vehicle is tracked with the constant-velocity motion model. The state and measurement equations of Kalman filter are defined in Eq. (2).

$$\begin{aligned} x_k &= Ax_{k-1} + w_{k-1} \\ z_k &= Hx_k + v_k \end{aligned} \quad (2)$$

In Eq. (2), x_{k-1} denotes the state at frame $k-1$, z_k denotes the measurement at frame k . The random variables w_k and v_k represent process and measurement noise. They are assumed to be independent, white, and normal distributed and defined as Eq. (3).

$$\begin{aligned} p(w) &\sim N(0, Q) \\ p(v) &\sim N(0, R) \end{aligned} \quad (3)$$

where, \mathbf{Q} denotes the process noise covariance matrix and \mathbf{R} denotes the measurement noise covariance matrix. In order to determine whether an observed vehicle belongs to a new incoming vehicle or a previously existed vehicle, we propose an efficient

matching algorithm to calculate the matching distance. For the texture matching, the texture feature can be generated from the canny edge [11] and the distance transform [19]. With the canny edge detection method we can retrieve the rough contours of the detected moving vehicles that is hard to be influenced by the illumination variation and shadow shown in Fig. 7. Then, the distant transform is applied to the canny edge context to retrieve the texture content for the target matching.

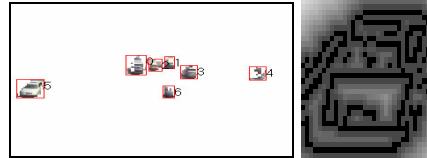


Fig. 7. The extractions of the texture features for each tracked vehicle

Distance transform matching is a technique for finding the matching distance between two different edge features by minimizing shifted template matching defined in Eq. (4).

$$\min_{q,r} \left(\sum_{q=0}^{M-O} \sum_{r=0}^{N-P} \left(\sum_{m=0}^M \sum_{n=0}^N [Tem(m,n) - Ref(o+q, p+r)] \right) \right), \quad (4)$$

where Tem is the template feature of size $M \times N$ and Ref is the reference feature of size $O \times P$. $Tem(x, y)$ denotes the pixel value in the template feature and $Ref(x, y)$ denotes the pixel value in the reference feature at the position (x, y) . If the matching value is less than the specified threshold, we can initialize a new Kalman filter to track or update the Kalman filter with observed vehicle position.

3.2 Vehicle Counting

Here, we set a virtual detecting region on the entire road to monitor whether the vehicles reach the detecting region or not. When a tracked vehicle touches the region, we start to monitor its trajectory within the detecting region. If the vehicle's trajectory

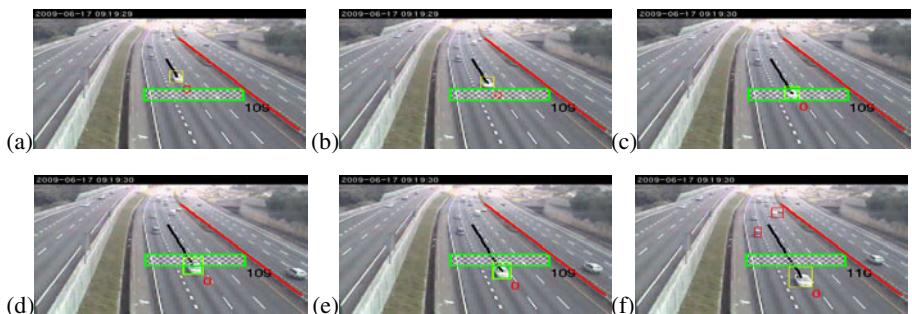


Fig. 8. The procedures of vehicle counting. The tracked vehicle is shown from (a) to (b). The tracked vehicle is driving through the detection region is shown from (c) to (e). (f) The tracked vehicle passes the detection region.

satisfies the following two conditions, then the vehicle is counted. First, the vehicle is detected in the virtual detecting region. Second, the length of the trajectory of a tracked vehicle must larger than a specified length threshold (100 pixels). Some examples of vehicle counting are shown in Fig. 8. In Fig. 8, the red line is used to separate the different directions of traffic flow and the green region is the detection region for counting. The black text represents the current number of passed vehicles.

4 Experimental Results

In this section, the video sequence of PetsD2TeC2 [20] and the real traffic videos captured from the web site of Taiwan Area National Freeway Bureau (<http://www.nfreeway.gov.tw>) are used as the test videos.

4.1 The Detection of Moving Vehicles

Here we use the PetsD2TeC2 video sequence to evaluate the performance of foreground detection for our method, traditional frame difference method, and the method of background subtracting with Gaussian Mixture Model method.

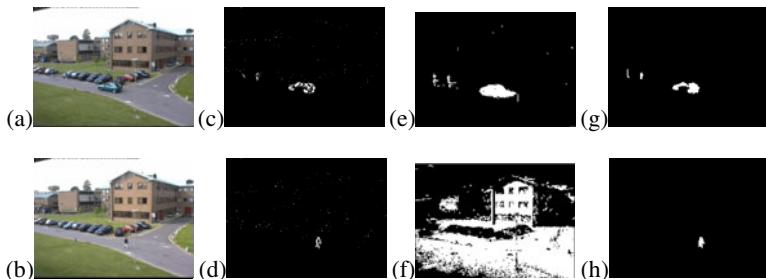


Fig. 9. The examples of foreground detection at frames #1023 and #2517. The original frames are shown from (a) to (b). The foreground detection using the conventional frame difference method is shown from (c) to (d). The foreground detection using the GMM background modeling method is shown from (e) to (f). The foreground detection using our proposed method is show from (g) to (h).

The block size is chosen as 2×2 to fit the size of moving object in our system. The lower bound of object size is set to be 10 blocks for the object detection. The foreground detections in Fig. 9-(c)(d) show the scene adaptability and fragmented region for the frame difference method. The foreground detections in Fig. 9-(e)(f) show the object completeness of the background modeling method [20]. Fig. 9-(f) shows a noisy image that is generated by the inefficient background updating process. The foreground detections in Fig. 9-(g)(h) show the scene adaptability and detection completeness of our proposed method. It is obvious that our method outperforms the other typical methods in terms of scene adaptability and detection completeness. Fig. 10-(a) shows the video sequence of traffic scene on the freeway. Fig. 10-(b) shows that the bounding boxes on the detected vehicles.

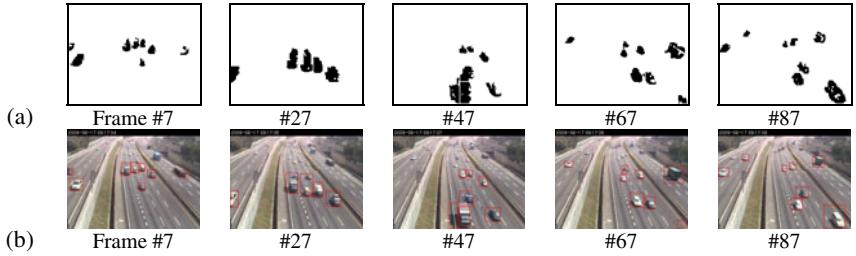


Fig. 10. Moving object detection. (a) A traffic flow scene. (b) The regions of precise moving objects are labeled with red rectangular box.

4.2 Vehicle Tracking and Counting

The vehicle tracking is illustrated in Fig. 11. In Fig. 11-(a), the moving vehicles start to be tracked. Fig. 11-(b)(c) show the correct labeling of the tracked vehicles after the merging and splitting of the two objects. The red rectangular represents the untracked object and the yellow rectangular represent the tracked vehicle. The red number denotes the number of the tracked vehicle.



Fig. 11. Vehicle tracking on crowded vehicle scene

After vehicle tracking, we set a detecting region on the roadway to count the moving vehicles. The detecting region across multiple lanes is used to count vehicles. The performance of vehicle counting is illustrated in Table 1. In the first and second video clips, the density of traffic flow is low. Hence, the performance of the vehicle tracking is satisfied. In the third video clip, the performance is reduced because a few buses introduce the some occlusions and wrong segmentation problems.

Table 1. The accuracy analysis of the vehicle counting

Heading level	1 st video clip	2 nd video clip	3 rd video clip
Actual number	85	88	76
Detected number	90	94	85
False positive	8	9	11
False negative	3	3	2
accuracy	89%	88%	83%

5 Conclusions

In this paper, we propose a novel vehicle detection method without background modeling in which the modified block-based frame differential method, the precise object

region extraction with dual foregrounds, the foreground segmentation, and the true object verification are integrated to develop a scene adaptive vehicle detection system. The texture-based target tracking method is proposed to track each detected target and then apply the virtual-loop detector to analyze the traffic flow. Experimental results show that our proposed system can work in real time with the rate above 20 fps and the accuracy of vehicle counting can approach 86%.

References

- [1] Dai, X., Khorram, S.: Performance of Optical Flow Techniques. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 236–242 (1992)
- [2] Kamijo, S., Matsushita, Y., Ikeuchi, K., Sakauchi, M.: Traffic Monitoring and Accident Detection at Intersections. *IEEE Transactions on Intelligent Transportation Systems* 1(2), 703–708 (2000)
- [3] Zhou, J., Gao, D., Zhang, D.: Moving Vehicle Detection for Automatic Traffic Monitoring. *IEEE Transactions On Vehicular Technology* 56(1), 51–59 (2007)
- [4] Lien, C.C., Wang, J.C., Jiang, Y.M.: Multi-Mode Target Tracking on a Crowd Scene. In: Proceedings of the Third International Conference on International Information Hiding and Multimedia Signal Processing, vol. 02, pp. 427–430 (2007)
- [5] Jain, R., Martin, W., Aggarwal, J.: Segmentation through the Detection of Changes due to Motion. *Compute Graph and Image Processing* 11, 13–34 (1979)
- [6] Wang, K.F., Li, Z.J., Yao, Q.M., Huang, W.L., Wang, F.Y.: An Automated Vehicle Counting System for Traffic Surveillance. In: IEEE International Conference on Vehicular Electronics and Safety (ICVES), pp. 1–6 (2007)
- [7] Lei, M., Lefloch, D., Gouton, P., Madani, K.: A Video-based Real-time Vehicle Counting System Using Adaptive Background Method. In: IEEE International Conference on Signal Image Technology and Internet Based Systems, pp. 523–528 (2008)
- [8] Qin, B., Zhang, M., Wang, S.: The Research on Vehicle Flow Detection in Complex Scenes. In: IEEE International Symposium on Information Science and Engineering, vol. 1, pp. 154–158 (2008)
- [9] Chen, T.H., Lin, Y.F., Chen, T.Y.: Intelligent Vehicle Counting Method Based on Blob Analysis in Traffic Surveillance. In: Second International Conference on Innovative Computing, Information and Control, pp. 238–238 (2007)
- [10] Baş, E., Tekalp, A.M., Salman, F.S.: Automatic Vehicle Counting from Video for Traffic Flow Analysis. In: IEEE Intelligent Vehicles Symposium, pp. 392–397 (2007)
- [11] Canny, J.: A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8(6), 679–698 (1986)
- [12] Liao, P.S., Chen, T.S., Chung, P.C.: A Fast Algorithm for Multilevel Thresholding. *Journal of Information Science and Engineering* 17, 713–727 (2001)
- [13] Dellepiane, S.G., Fontana, F., Vemazza, G.L.: Nonlinear Image Labeling for Multivalued Segmentation. *IEEE Transactions on Image Processing* 5(3), 429–446 (1996)
- [14] Koga, T., Iinuma, K., Hirano, A., Iijima, Y., Ishiguro, T.: Motion Compensated Interframe Coding for Video Conferencing. In: Proc. Nat. Telecommunication Conf., pp. G5.3.1–G5.3.5 (1981)
- [15] Wren, C., Azarbayejani, A., Darrell, T., Pentland, A.: Pfnder: Real-time Tracking of the Human Body. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(7), 780–785 (1997)

- [16] Xu, M., Orwell, J., Lowey, L., Thirde, D.: Architecture and Algorithms for Tracking Football Players with Multiple Cameras. In: IEEE Proceedings Vision, Image and Signal Processing, vol. 152(2), pp. 232–241 (2005)
- [17] Welch, G., Bishop, G.: An Introduction to the Kalman Filter (2004)
- [18] Nummiaro, K., Meier, E.K., Gool, L.J.V.: An Adaptive Color-based Particle Filter. Image Vision Computing 21(1), 99–110 (2002)
- [19] Ghafoor, A., Iqbal, R.N., Khan, S.: Image Matching Using Distance Transform. In: Bigun, J., Gustavsson, T. (eds.) SCIA 2003. LNCS, vol. 2749, pp. 654–660. Springer, Heidelberg (2003)
- [20] Huang, T., Qiu, J., Sakayori, T., Goto, S., Ikenaga, T.: Motion Detection Based On Background Modeling And Performance Analysis For Outdoor Surveillance. In: IEEE International Conference on Computer Modeling and Simulation, pp. 38–42 (2009)

Effective Color-Difference-Based Interpolation Algorithm for CFA Image Demosaicking

Yea-Shuan Huang and Sheng-Yi Cheng

Department of Computer Science Information Engineering, Chung Hwa University, 707, Sec.2,
WuFu Rd., Hsinchu, Taiwan 300, R.O.C.

Abstract. This paper proposes an effective color-difference-based (ECDB) interpolation algorithm for CFA Image demosaicking. A CFA image consists of a set of spectrally selective filters which are arranged in an interleaved pattern such that only one of color component is sampled at each pixel location. To improve the quality of reconstructed full-color images from color filter array (CFA) images, the ECDB algorithm first analyzes the neighboring samples around a green missing pixel to determine suitable samples for interpolating the value of this green missing pixel. After finishing the interpolation operations of all the green missing pixels, a complete green plane (i.e \bar{G} plane) can be obtained. The ECDB algorithm then makes use of the high correlation between R, G, and B planes to produce the red–green and blue–green color difference planes and further reconstructs the red and blue planes in successive operations. Because of the green plane provides twice information than red and blue planes, the algorithm exploits the information of green plane more than that of red/blue plane so that the full color image can be reconstructed more accurately. In essence, the ECDB algorithm uses the red–green and blue–green color difference planes, and develops different conditional operations according to the horizontal, vertical, and diagonal neighboring pixel information with suitable weighting technique. The experimental results demonstrate that the proposed algorithm has outstanding performance.

Keywords: Color filter array, Demosaicking, Interpolation.

1 Introduction

In nowadays, Most of electronic devices such as digital still cameras, mobile phones, and PDAs use a single image sensor to capture digital images, which usually consist of three color components (red, green, and blue) at each pixel location. However, the surface of single image sensor is covered with a color filter array (CFA). A CFA image consists of a set of spectrally selective filters which are arranged in an interleaved pattern such that only one of color component is sampled at each pixel location. In order to render a full color image, image interpolation algorithm is performed to estimate the other two missing color components and the image interpolation is called as CFA demosaicking algorithm (spectral interpolation).

To obtain a good demosaicked color image, a lot of demosaicking algorithms has been proposed, such as bilinear interpolation, cubic spline interpolation, nearest-neighbor replication, etc. Although, these single plane interpolation methods are easy to implement, they produce severe color artifacts around sharp areas. Recently,

another method called edge-sensing correlation-correction (ESCC) has been proposed. It firstly interpolates green missing pixels using eight neighboring green samples with weighted technique, which considers the edges in all possible directions, i.e. diagonal, horizontal, and vertical. Next, it interpolates the missing red/ blue pixels by exploiting inter-plane color difference (red minus green/ blue minus green), which assumes that the hue does not abruptly change in a local area. The ESCC can produce pleasing demosaicked images to human vision in most of cases. However, there are considerable false colors occurred at thin line areas of demosaicked image because the ESCC uses unsuitable samples to estimate the color values of missing pixels.

In this paper, we propose an effective demosaicking algorithm to reconstruct full-color images from CFA images based on Bayer CFA pattern, and the reconstructed image can provide very good visual quality. The proposed demosaicking algorithm is a new kind of effective color-difference-based interpolation (ECDB) demosaicking algorithm. The ECDB algorithm starts from green plane to produce reconstructed full-color image, because of the green plane provides twice information than red and blue planes. Therefore, When interpolating the green missing pixels, the ECDB algorithm will make use of neighboring samples to analyze the edge direction (vertical, horizontal or diagonal), then according to specially designed operations to select suitable samples for estimating the green missing pixel. Once the green plane is produced, the ECDB algorithm exploits both the demosaicked green plane and the original red/blue plane to produce red-green and blue-green color different planes, and then the color different planes are interpolated by using neighboring samples with suitable weighting technique. Finally, by using the green plane, red-green and blue-green color difference planes high-quality demosaicked full-color images can be reconstructed.

2 The Proposed Scheme

2.1 The Demosaicking Procedure of Green Plane

Figure 1 shows the flowchart of the proposed ECDB algorithm. The ECDB algorithm starts with the procedures of recovering a CFA image with the demosaicked green plane. Then it exploits the R-G and B-G color difference planes to reconstruct the red and blue planes, respectively.

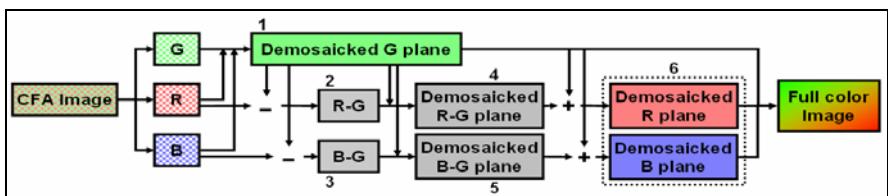


Fig. 1. The flowchart of the ECDB algorithm

During interpolating the green plane, we will take consideration for the neighboring pixels around a currently processing pixel. In general, high continuity neighboring pixels will have higher attribute (such as color or texture) consistency than low continuity neighboring pixels. To interpolate the missing color value of a pixel A we should

choose only those pixels having high attribute consistency from all the neighboring pixels of A and abandon those neighboring pixels having low attribute consistency. Centered with pixel A, a region E of 7×7 pixels is defined. For each pixel A, we use the edge strength to measure the continuity and consistency among its neighboring pixels, and also use the edge strengths of different directions to determine the consistency direction (horizontal or vertical) of its corresponding region E. Obviously, the direction having the weaker edge strength probably corresponds to the consistency direction. If both vertical and horizontal edge strengths are very weak, then the region will be determined to be a smooth region. On the contrary, if both vertical and horizontal edge strengths are strong, then the region will be determined to be a complex region. Conceptually to interpolate the missing color value of pixel A, only those pixels along the consistency direction will be used. However, even along the consistency direction probably only a part of pixels are truly consistent and the other are not. Therefore, we further judge the region E into five different cases so that more appropriate pixels can be chosen. Case 1, the consistency direction is horizontal but only the right or left half of E is consistent. Case 2, the consistency direction is vertical but only the upper or lower half of E is consistent. Case 3, the consistency direction is horizontal and both the right and left half of E are consistent. Case 4, the consistency direction is vertical and both the upper and lower half of E are consistent. Case 5, both horizontal and vertical direction are either both consistent or both not consistent. With the above judgment, we design the following mechanism to interpolate the missing green color. In the designed mechanism, we also developed a novel weight technique to give each chosen pixel sample a suitable weight value for calculation.

Figure 2. displays the configuration of related pixels for estimating the green value of pixel located at “c1”. Because Bayer pattern have only one color value at each pixel location, two color values of each pixel are missing and they need to be estimated. In this section, we introduce how to interpolate green color value on the blue plane in detail, and it will be similar to interpolate the green color value on the red plane.

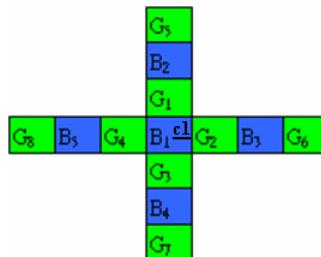


Fig. 2. The configuration of related pixels for estimating the green value of pixel located at “c1”

2.2 Principle of Variable Naming

In the proposed algorithm we use quite a few variables. In order to let variables can be understood more easily, we introduce the variable naming principle in this section. When we define a variable, usually it contains an abbreviated prefix, superscript and suffix. For prefix, ‘e’ denotes “edge strength”, ‘w’ denotes “weight”, and ‘d’ denotes “demosaicked value”. For superscript, ‘r’, ‘g’, and ‘b’ denote the red color, the green

color and the blue color, respectively. Also, ‘r-g’ denotes the color difference of the red color and the green color, ‘b-g’ denotes the color difference of the blue color and the green color, ‘gb’ denotes in the green color channel reference to the blue color value, ‘rb’ denotes in the color difference ‘r-g’ channel reference to the blue color value, ‘rg’ denotes in the color difference ‘r-g’ channel reference to the green color value, and ‘a’ denotes “adaptive” to represent more than one color channel are used to interpolate a missing value. For suffix, only the direction attribute, weight attribute, and the summation attribute will appear. In suffix, symbol ‘w’ denotes the weight attribute, symbol ‘s’ denotes the summation attribute, and there are 12 direction-related attributes which are listed in Table 1 with their names and their corresponding symbols. For example, variable e_1^g denotes the vertical edge strength computed from the green color channel.

Table 1. List of suffix attribute and symbol

Suffix attribute	symbol	Suffix attribute	symbol	Suffix attribute	symbol
Vertical		Top	↑	Top right	→↑
Horizontal	—	Bottom	↓	Top left	←↑
Diagonal	/	Left	←	Bottom right	→↓
Reverse diagonal	\	Right	→	Bottom left	←↓

If the estimated value on the horizontal direction has high attribute consistence for both side (left side and right side), we give an equal weight value to both side; otherwise, we give different weight values to both sides. The same concept can be applied on vertical direction. Let e_1^{gb} , e_{\uparrow}^g and e_{\downarrow}^g be the blue color vertical edge strength, green color upper edge strength and green color lower edge strength of this configuration, respectively. Let e_{\rightarrow}^{gb} , e_{\rightarrow}^g and e_{\leftarrow}^g be the blue color horizontal edge strength, green color right edge strength and green color left edge strength of this configuration, respectively. The values of e_1^{gb} , e_{\uparrow}^g , e_{\downarrow}^g , e_{\rightarrow}^{gb} , e_{\rightarrow}^g and e_{\leftarrow}^g are computed by Eq. 2-1 as

$$\begin{aligned} e_1^{gb} &= |2*B_1 - B_2 - B_4|, & e_{\rightarrow}^{gb} &= |2*B_1 - B_3 - B_5|, \\ e_{\uparrow}^g &= |G_1 - G_3| + 2*|G_1 - G_5|, & e_{\rightarrow}^g &= |G_2 - G_4| + 2*|G_2 - G_6|, \\ e_{\downarrow}^g &= |G_1 - G_3| + 2*|G_3 - G_7|, & e_{\leftarrow}^g &= |G_2 - G_4| + 2*|G_4 - G_8|. \end{aligned} \quad (2-1)$$

Also, let e_1^g and e_{\rightarrow}^g be the edge strength values of the vertical and horizontal directions of this configuration, respectively. The value of e_1^g is according to the calculation of e_1^{gb} , e_{\uparrow}^g and e_{\downarrow}^g , and the value of e_{\rightarrow}^g is according to the calculation of e_{\rightarrow}^{gb} , e_{\rightarrow}^g and e_{\leftarrow}^g . The above two edge strengths can be computed by Eq. 2-2 as

$$e_1^g = 1 + e_1^{gb} + \frac{e_{\uparrow}^g + e_{\downarrow}^g}{2}, \quad e_{\rightarrow}^g = 1 + e_{\rightarrow}^{gb} + \frac{e_{\rightarrow}^g + e_{\leftarrow}^g}{2}. \quad (2-2)$$

Let d_{\rightarrow}^b , d_{\rightarrow}^g and d_{\leftarrow}^g be three estimated horizontal-direction variables which are used to calculate the estimated missing green value \hat{d}_{\perp} , w_{\rightarrow}^g and w_{\leftarrow}^g be the weight values

of d_{\rightarrow}^b and d_{\leftarrow}^b , d_{\downarrow}^g be the estimated values of vertical direction, d_{\uparrow}^b be the difference of B_1 and B_2 , d_{\downarrow}^b be the difference of B_1 and B_4 , w_{\uparrow}^g and w_{\downarrow}^g be the weight values of d_{\uparrow}^b and d_{\downarrow}^b . The 10 parameters (d_{-}^g , d_{\rightarrow}^b , d_{\leftarrow}^b , w_{\leftarrow}^g , w_{\rightarrow}^g , d_{\downarrow}^g , d_{\uparrow}^b , d_{\downarrow}^b , w_{\uparrow}^g and w_{\downarrow}^g) are calculated as bellow:

$$d_{-}^g = \frac{G_2 + G_4}{2}, \quad d_{\rightarrow}^b = \frac{B_1 - B_3}{4}, \quad d_{\leftarrow}^b = \frac{B_1 - B_5}{4}, \quad d_{\downarrow}^g = \frac{G_1 + G_3}{2}, \quad d_{\uparrow}^b = \frac{B_1 - B_2}{4}, \quad d_{\downarrow}^b = \frac{B_1 - B_4}{4} \quad (2-3)$$

$$w_{\leftarrow}^g = \frac{e_{\rightarrow}^g}{e_{\rightarrow}^g + e_{\leftarrow}^g}, \quad w_{\rightarrow}^g = 1 - w_{\leftarrow}^g, \quad w_{\uparrow}^g = \frac{e_{\downarrow}^g}{e_{\uparrow}^g + e_{\downarrow}^g}, \quad w_{\downarrow}^g = 1 - w_{\uparrow}^g \quad (2-4)$$

Let d_{w-}^a and $d_{w\downarrow}^a$ be the estimated values of horizontal and vertical directions with weight, respectively. d_{-}^a and d_{\downarrow}^a are the estimated values of horizontal and vertical directions without weight, respectively. It can be calculated as bellow:

$$\begin{aligned} d_{w-}^a &= d_{-}^g + w_{\leftarrow}^g \times d_{\leftarrow}^b + w_{\rightarrow}^g \times d_{\rightarrow}^b, \quad d_{w\downarrow}^a = d_{\downarrow}^g + w_{\uparrow}^g \times d_{\uparrow}^b + w_{\downarrow}^g \times d_{\downarrow}^b, \\ d_{-}^a &= d_{-}^g + d_{\rightarrow}^b + d_{\leftarrow}^b, \quad d_{\downarrow}^a = d_{\downarrow}^g + d_{\uparrow}^b + d_{\downarrow}^b. \end{aligned} \quad (2-5)$$

Then, the central green missing sample “c1” can be estimated by the following Equation:

$$\hat{d}_{\text{c1}}^g = \begin{cases} d_{w-}^a & , \text{ if } e_{\downarrow}^g / e_{\leftarrow}^g > R \text{ and } (e_{\downarrow}^g + e_{\leftarrow}^g) > T_1 \text{ and } e_{\downarrow}^g < T_2 \text{ and } |e_{\rightarrow}^g - e_{\leftarrow}^g| > D_{eg}; \\ d_{w\downarrow}^a & , \text{ if } e_{\downarrow}^g / e_{\leftarrow}^g > R \text{ and } (e_{\downarrow}^g + e_{\leftarrow}^g) > T_1 \text{ and } e_{\downarrow}^g < T_2 \text{ and } |e_{\uparrow}^g - e_{\downarrow}^g| > D_{eg}; \\ d_{-}^a & , \text{ if } e_{\downarrow}^g / e_{\leftarrow}^g > R \text{ and } (e_{\downarrow}^g + e_{\leftarrow}^g) > T_1 \text{ and } e_{\downarrow}^g < T_2 \text{ and } |e_{\rightarrow}^g - e_{\leftarrow}^g| \leq D_{eg}; \\ d_{\downarrow}^a & , \text{ if } e_{\downarrow}^g / e_{\leftarrow}^g > R \text{ and } (e_{\downarrow}^g + e_{\leftarrow}^g) > T_1 \text{ and } e_{\downarrow}^g < T_2 \text{ and } |e_{\uparrow}^g - e_{\downarrow}^g| \leq D_{eg}; \\ d_{\downarrow}^a \times w_{\downarrow}^g + d_{-}^a \times w_{-}^g & , \text{ otherwise}. \end{cases} \quad (2-6)$$

where \hat{d}_{c1}^g is the estimated value for the green missing pixel on the red or blue sample location at “c1”. In Eq. 2-6, R , T_1 , T_2 and D_{eg} are four threshold values. R is the threshold for the ratio of e_{\downarrow}^g and e_{\leftarrow}^g . If the value of $e_{\downarrow}^g / e_{\leftarrow}^g$ is larger than R , it means the configuration of “c1” has relatively high continuity in the horizontal direction, else if the value of $e_{\downarrow}^g / e_{\leftarrow}^g$ is larger than R , then it means the configuration of “c1” has relatively high continuity on the vertical direction, otherwise, there is no obvious horizontal and vertical continuity at position “c1”. If the edge strength $|e_{\downarrow}^g + e_{\leftarrow}^g|$ is larger than T_1 , it means the central pixel at “c1” position is located at a non-smooth area with low continuity in either horizontal or vertical direction; otherwise pixel “c1” is located at a smooth area with high continuity in both horizontal and vertical directions. If the edge strength (either e_{\downarrow}^g or e_{\leftarrow}^g) is less than T_2 , it means pixel “c1” is located at a non-complex area with high continuity in either horizontal or vertical direction; otherwise, “c1” is located at a complex area with low continuity in both

horizontal and vertical direction. e_{\rightarrow}^g and e_{\leftarrow}^g are the edge strengths of the horizontal direction for the right side and the left side at “c1”, respectively. e_{\uparrow}^g and e_{\downarrow}^g are the edge strengths of the vertical direction for the top side and the bottom side, respectively. If the value of $|e_{\rightarrow}^g - e_{\leftarrow}^g|$ is larger than D_{eg} , it means the edge strengths of the right and the left are significantly different, otherwise, the two values have no large difference. For the vertical side, the same principle can be applied to decide whether the edge strengths of the top side and the bottom side are different enough by comparing $|e_{\uparrow}^g - e_{\downarrow}^g|$ with D_{eg} .

In Eq. 2-6, there are five situations and each situation has its own method to select the neighboring pixels for interpolating the missing green value. The first situation corresponds to a horizontally half-side uniform configuration which satisfies the following four conditions: (1) the horizontal edge strength should be relatively smaller than the vertical edge strength (i.e. $e_{\uparrow}^g / e_{\downarrow}^g > R$), (2) the total horizontal and vertical edge strength should be not too small (i.e. $e_{\uparrow}^g + e_{\downarrow}^g > T_1$), (3) the horizontal edge strength should be small enough (i.e. $e_{\rightarrow}^g < T_2$), and (4) only half of the horizontal direction (either the right-half or the left-half) is uniform and the other half is non-uniform (i.e. $|e_{\rightarrow}^g - e_{\leftarrow}^g| > D_{eg}$). If all above four conditions are satisfied, then we estimate the missing value \hat{d}^g by the value of d_{w-}^a .

The second, third, and fourth situations correspond to a vertically half-side uniform configuration, a horizontally two-side uniform configuration, and a vertically two-side uniform configuration, respectively. If the green missing pixel does not belong to the above four situations, it then belongs to the fifth situation which corresponds to either a non-uniform or a uniform configuration in both horizontal and vertical directions. Readers can easily understand the required conditions of each situation because they are similar to the condition illustration of the first situation. In the fifth situation we will base on the individual green color continuity in the respectively vertical and horizontal direction to give each direction the suitable weight. So that we interpolate the central green missing value according to d_{-}^a, d_{\uparrow}^a and their weights w_{\uparrow}^g and w_{-}^g by the following Equation.

$$w_{\uparrow}^g = \frac{e_{\downarrow}^g}{e_{\uparrow}^g + e_{\downarrow}^g}, \quad w_{-}^g = 1 - w_{\uparrow}^g. \quad (2-7)$$

2.3 The Demosaicking Procedure of R-G/B-G Color Difference Planes

Let $\mathbf{R-G}$ denote the color difference plane of the red plane and the demosaicked green plane and $\mathbf{B-G}$ denote the color difference plane of the blue plane and the demosaicked green plane. That is $\mathbf{R-G} = \mathbf{R} - \overline{\mathbf{G}}$ and $\mathbf{B-G} = \mathbf{B} - \overline{\mathbf{G}}$. Both $\mathbf{R-G}$ and $\mathbf{B-G}$ inherently are low-pass signal which can be utilized to reduce the estimated errors of the demosaicked images, and the ECDB algorithm uses them to reconstruct the red and blue planes, individually. Since both $\mathbf{R-G}$ and $\mathbf{B-G}$ are derived by similar operations, only the derivation of $\mathbf{R-G}$ is described here. Because there are many red-missing pixels in the red plane R, many pixels will accordingly have no values in the $\mathbf{R-G}$ plane. These

pixels are called the “missing **R-G** pixels” which can be categorized into three classes: diagonal class, vertical class, and horizontal class. The class criteria are (1) if its left and right **R-G** pixels have values, it is attributed to the horizontal class; (2) if its top and bottom **R-G** pixels have values, it is attributed to the vertical class; (3) otherwise, it is attributed to the diagonal class. Because in our design when reconstructing a **R-G** pixel value of an either horizontal-class or vertical-class pixel, the values of some pixels belonging to diagonal class will be referenced; but the **R-G** pixel value of a diagonal-class pixel can be estimated directly without referencing the values of horizontal-class and vertical-class pixels. Therefore, the diagonal-class pixels should be estimated first, and the pixels of the other two classes are estimated later.

2.4 Value Estimation of the Missing Diagonal-Class R-G Pixels

In order to describe the estimated value of a missing **R-G** pixel, Figure 3 is provided to show the geometric relations of the missing **R-G** pixels. For estimating the **R-G** value of B_1 , let e_{\perp}^a , e_{-}^a , $e_{/}^a$ and e_{\backslash}^a be the edge strengths in the vertical, horizontal, diagonal and reverse diagonal directions, respectively. The four edge strengths ($e_i^a, \forall i \in \{\perp, -, /, \backslash\}$) are defined as

$$e_{\perp}^a = e_{\perp}^{rg} + e_{\perp}^{rb}, \quad e_{/}^a = \frac{e_{\rightarrow\uparrow}^r + e_{\leftarrow\downarrow}^r}{2} + e_{/}^{rb}, \quad e_{-}^a = e_{-}^{rg} + e_{-}^{rb}, \quad e_{\backslash}^a = \frac{e_{\leftarrow\uparrow}^r + e_{\rightarrow\downarrow}^r}{2} + e_{\backslash}^{rb}. \quad (2-8)$$

$$e_{\perp}^{rb} = |2 * B_1 - B_6 - B_2|, \quad e_{/}^{rb} = |2 * B_1 - B_4 - B_8|, \quad e_{-}^{rb} = |2 * B_1 - B_7 - B_3|, \quad e_{\backslash}^{rb} = |2 * B_1 - B_5 - B_9|,$$

$$e_{\perp}^{rg} = |G_{13} - G_1| + |G_1 - G_3| + |G_3 - G_{19}|, \quad e_{-}^{rg} = |G_{22} - G_4| + |G_4 - G_2| + |G_2 - G_{16}|.$$

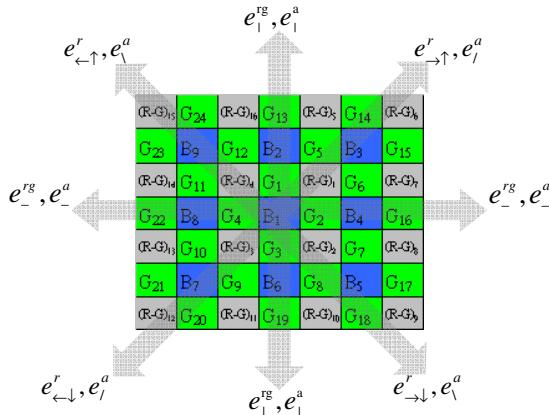


Fig. 3. The to-be-estimated **R-G** value on the blue sample B_1

Let $e_{\rightarrow\uparrow}^r$, $e_{\leftarrow\downarrow}^r$, $e_{\leftarrow\uparrow}^r$ and $e_{\rightarrow\downarrow}^r$ be the edge strengths of the missing **R-G** pixels on the top-right, the bottom-left, the top-left, and the bottom-right direction, respectively. The values of $e_{\rightarrow\uparrow}^r$, $e_{\leftarrow\downarrow}^r$, $e_{\leftarrow\uparrow}^r$ and $e_{\rightarrow\downarrow}^r$ can be given by

$$\begin{aligned} e_{\rightarrow \uparrow}^r &= |R_3 - R_1| + 2 * |R_1 - R_6|, & e_{\leftarrow \uparrow}^r &= |R_4 - R_2| + 2 * |R_4 - R_{15}|, \\ e_{\leftarrow \downarrow}^r &= |R_3 - R_1| + 2 * |R_3 - R_{12}|, & e_{\rightarrow \downarrow}^r &= |R_4 - R_2| + 2 * |R_2 - R_9|. \end{aligned} \quad (2-9)$$

Let $(\mathbf{R}\text{-}\mathbf{G})_1$, $(\mathbf{R}\text{-}\mathbf{G})_2$, $(\mathbf{R}\text{-}\mathbf{G})_3$ and $(\mathbf{R}\text{-}\mathbf{G})_4$ be the values of the diagonal and reverse diagonal $\mathbf{R}\text{-}\mathbf{G}$ pixels around B_1 , d_i^{r-g} be the average of $(\mathbf{R}\text{-}\mathbf{G})_1$ and $(\mathbf{R}\text{-}\mathbf{G})_3$, d_{\backslash}^{r-g} be the average of $(\mathbf{R}\text{-}\mathbf{G})_2$ and $(\mathbf{R}\text{-}\mathbf{G})_4$. $w_{\rightarrow \uparrow}^r$, $w_{\leftarrow \downarrow}^r$, $w_{\leftarrow \uparrow}^r$ and $w_{\rightarrow \downarrow}^r$ be the weight of $(\mathbf{R}\text{-}\mathbf{G})_1$, $(\mathbf{R}\text{-}\mathbf{G})_2$, $(\mathbf{R}\text{-}\mathbf{G})_3$ and $(\mathbf{R}\text{-}\mathbf{G})_4$, respectively. These 6 parameters (d_i^{r-g} , d_{\backslash}^{r-g} , $w_{\rightarrow \uparrow}^r$, $w_{\leftarrow \downarrow}^r$, $w_{\leftarrow \uparrow}^r$ and $w_{\rightarrow \downarrow}^r$) are calculated by

$$\begin{aligned} d_i^{r-g} &= \frac{(R-G)_1 + (R-G)_3}{2}, & d_{\backslash}^{r-g} &= \frac{(R-G)_2 + (R-G)_4}{2}. \\ w_{\rightarrow \uparrow}^r &= \frac{e_{\leftarrow \downarrow}^r}{e_{\rightarrow \uparrow}^r + e_{\leftarrow \downarrow}^r}, & w_{\leftarrow \downarrow}^r &= 1 - w_{\leftarrow \downarrow}^r, & w_{\leftarrow \uparrow}^r &= \frac{e_{\rightarrow \downarrow}^r}{e_{\leftarrow \uparrow}^r + e_{\rightarrow \downarrow}^r}, & w_{\rightarrow \downarrow}^r &= 1 - w_{\rightarrow \downarrow}^r. \end{aligned} \quad (2-10)$$

If the estimated value on the diagonal direction has high attribute consistency for both side (top-right and bottom-left), we give an equal weight value to both side (i.e. d_i^{r-g}); otherwise, we give different weight values to both sides (i.e. $w_{\rightarrow \uparrow}^r \times (\mathbf{R}\text{-}\mathbf{G})_1 + w_{\leftarrow \downarrow}^r \times (\mathbf{R}\text{-}\mathbf{G})_3$). If the estimated value on the reverse diagonal direction has high attribute consistency for both side (top-left and bottom-right), we give an equal weight value to both side (i.e. d_{\backslash}^{r-g}); otherwise, we give different weight values to both sides (i.e. $w_{\leftarrow \uparrow}^r \times (\mathbf{R}\text{-}\mathbf{G})_4 + w_{\rightarrow \downarrow}^r \times (\mathbf{R}\text{-}\mathbf{G})_2$). When there is no significant consistency direction either on diagonal or reverse diagonal direction. We will base on the individual R-G plane continuity in the respectively top-right, bottom-left, top-left and bottom-right direction to give each direction the suitable weight values. So that we interpolate the missing R-G value according to $(\mathbf{R}\text{-}\mathbf{G})_1$, $(\mathbf{R}\text{-}\mathbf{G})_2$, $(\mathbf{R}\text{-}\mathbf{G})_3$ and $(\mathbf{R}\text{-}\mathbf{G})_4$ and their weights $w_{\rightarrow \uparrow}^a$, $w_{\rightarrow \downarrow}^a$, $w_{\leftarrow \downarrow}^a$ and $w_{\leftarrow \uparrow}^a$ can given by

$$\begin{aligned} w_{\rightarrow \uparrow}^a &= \frac{1}{1 + |B_1 - B_3| + |G_1 - G_5| + |G_2 - G_6|}, & w_{\leftarrow \downarrow}^a &= \frac{1}{1 + |B_1 - B_7| + |G_3 - G_9| + |G_4 - G_{10}|}, \\ w_{\rightarrow \downarrow}^a &= \frac{1}{1 + |B_1 - B_5| + |G_2 - G_7| + |G_3 - G_8|}, & w_{\leftarrow \uparrow}^a &= \frac{1}{1 + |B_1 - B_9| + |G_4 - G_{11}| + |G_1 - G_{12}|}. \end{aligned} \quad (2-11)$$

The estimated \hat{d}_i^{rg} value of a diagonal-class missing pixel (B_1) becomes

$$\hat{d}_i^{r-g} = \begin{cases} w_{\rightarrow \uparrow}^r \times (\mathbf{R}\text{-}\mathbf{G})_1 + w_{\leftarrow \downarrow}^r \times (\mathbf{R}\text{-}\mathbf{G})_3, & \text{if } |e_{\rightarrow \uparrow}^r - e_{\leftarrow \downarrow}^r| \geq T_d \text{ and } |e_{\rightarrow \uparrow}^r + e_{\leftarrow \downarrow}^r| < T_2 \text{ and } \max_{i \in \{1, \dots, N\}} e_i^a / e_i^r > T_d \text{ and } e_i^a = \min_{i \in \{1, \dots, N\}} e_i^a; \\ w_{\leftarrow \uparrow}^r \times (\mathbf{R}\text{-}\mathbf{G})_4 + w_{\rightarrow \downarrow}^r \times (\mathbf{R}\text{-}\mathbf{G})_2, & \text{if } |e_{\leftarrow \uparrow}^r - e_{\rightarrow \downarrow}^r| \geq T_d \text{ and } |e_{\leftarrow \uparrow}^r + e_{\rightarrow \downarrow}^r| < T_2 \text{ and } \max_{i \in \{1, \dots, N\}} e_i^a / e_i^r > T_d \text{ and } e_i^a = \min_{i \in \{1, \dots, N\}} e_i^a; \\ d_i^{r-g}, & \text{if } \max_{i \in \{1, \dots, N\}} e_i^a / e_i^r > T_d \text{ and } e_i^a = \min_{i \in \{1, \dots, N\}} e_i^a; \\ d_{\backslash}^{r-g}, & \text{if } \max_{i \in \{1, \dots, N\}} e_i^a / e_i^r > T_d \text{ and } e_i^a = \min_{i \in \{1, \dots, N\}} e_i^a; \\ d_s^{r-g}, & \text{otherwise.} \end{cases} \quad (2-12)$$

In Eq. 2-12, T_d and T_2 are two threshold values. If the edge strength $|e_{\rightarrow \uparrow}^r - e_{\leftarrow \downarrow}^r| \geq T_d$, it means the edge strengths of the top-right and the bottom-left directions have a significant difference. If the edge strength $|e_{\rightarrow \uparrow}^r + e_{\leftarrow \downarrow}^r| < T_2$, it means the missing $\mathbf{R}\text{-}\mathbf{G}$ pixel

is located at a non-complex area. $\max_{\forall i \in \{1, -, /, \backslash\}} e_i^a$ and $\min_{\forall i \in \{1, -, /, \backslash\}} e_i^a$ denote the maximum and the minimum of e_1^a , e_-^a , $e/_^a$ and e/\backslash^a , respectively. $\max_{\forall i \in \{1, -, /, \backslash\}} e_i^a / e_i^a > T_d$ and $e/\backslash^a = \min_{\forall i \in \{1, -, /, \backslash\}} e_i^a$ means the region around B_1 has a high continuity along the diagonal direction and it is a non-complex area. Similarly, $\max_{\forall i \in \{1, -, /, \backslash\}} e_i^a / e_i^a > T_d$ and $e\backslash/_^a = \min_{\forall i \in \{1, -, /, \backslash\}} e_i^a$ means the region around B_1 has a high continuity along the reverse diagonal direction and it is a non-complex area.

In Eq. 2-12, there are five situations and each situation has its own method to select the neighboring pixels for interpolating the missing R-G value. The first situation corresponds to a diagonally half-side uniform configuration which satisfies the following four conditions: (1) only half of the diagonal direction (either the top-right-half or the bottom-left-half) is uniform and the other half is non-uniform (i.e. $|e_{\rightarrow\uparrow}^r - e_{\leftarrow\downarrow}^r| >= T_d$), (2) the diagonal edge strength should be small enough (i.e. $|e_{\rightarrow\uparrow}^r + e_{\leftarrow\downarrow}^r| < T_2$), (3) the diagonal edge strength should be relatively smaller than the maximum strengths (i.e. $\max_{\forall i \in \{1, -, /, \backslash\}} e_i^a / e_i^a > T_d$), (4) the diagonal edge strength should be the smallest among the four direction edge strengths (i.e. $e/\backslash^a = \min_{\forall i \in \{1, -, /, \backslash\}} e_i^a$). If all above four conditions are satisfied, then we estimate the missing R-G value by $w'_{\rightarrow\uparrow} \times (R-G)_1 + w'_{\leftarrow\downarrow} \times (R-G)_3$. The second, third, and fourth situations correspond to a reverse diagonally half-side uniform configuration, a diagonally two-side uniform configuration, and a reverse diagonally two-side uniform configuration, respectively. If the missing R-G pixel does not belong to the above four situations, it then belongs to the fifth situation which corresponds to either a non-uniform or a uniform configuration in both diagonal and reverse diagonal directions. Similar to the explanation of the first situation, readers can easily understand the required conditions of other situations. In the fifth situation we will base on the individual R-G pixel continuity in the respectively diagonal and reverse diagonal direction to give each direction the suitable weight. So that we interpolate the missing R-G value according to $(R-G)_1$, $(R-G)_2$, $(R-G)_3$ and $(R-G)_4$ and their corresponding weights $w_{\rightarrow\uparrow}^a$, $w_{\leftarrow\downarrow}^a$, $w_{\leftarrow\downarrow}^a$ and $w_{\leftarrow\uparrow}^a$ by the following Equation as

$$d_s^{r-g} = \frac{(R-G)_1 \times w_{\rightarrow\uparrow}^a + (R-G)_2 \times w_{\rightarrow\downarrow}^a + (R-G)_3 \times w_{\leftarrow\downarrow}^a + (R-G)_4 \times w_{\leftarrow\uparrow}^a}{\sum_{\forall j \in \{(\rightarrow\uparrow), (\leftarrow\downarrow), (\leftarrow\uparrow), (\rightarrow\downarrow)\}} w_j^a}. \quad (2-13)$$

where the d_s^{r-g} is the ratio of the weighted summation of $(R-G)_1$, $(R-G)_2$, $(R-G)_3$ and $(R-G)_4$ and the summation of $w_{\rightarrow\uparrow}^a$, $w_{\rightarrow\downarrow}^a$, $w_{\leftarrow\downarrow}^a$ and $w_{\leftarrow\uparrow}^a$.

2.5 Value Estimation of the Missing Vertical-Class R-G Pixels

Let w_1^{r-g} and w_-^{r-g} be the weights associated with the vertical and horizontal directions, respectively, and they are defined as

$$\begin{aligned}
 e_{-}^{r-g} &= |2*(R-G)_4 - (R-G)_8 - (R-G)_2| + |2*(R-G)_2 - (R-G)_4 - (R-G)_6|, \\
 e_{\perp}^{r-g} &= |2*(R-G)_1 - (R-G)_5 - (R-G)_3| + |2*(R-G)_3 - (R-G)_1 - (R-G)_7|. \\
 w_{\perp}^{r-g} &= 0.8 \times \frac{e_{\perp}^{r-g}}{e_{\perp}^{r-g} + e_{-}^{r-g}}, \quad w_{-}^{r-g} = 1 - w_{\perp}^{r-g}.
 \end{aligned} \tag{2-14}$$

where e_{-}^{r-g} and e_{\perp}^{r-g} denote the individual edge strengths of the horizontal and vertical directions. The larger e_{-}^{r-g} compares to e_{\perp}^{r-g} , the larger w_{\perp}^{r-g} will be. In figure 4, the R-G values of $(R-G)_1$, $(R-G)_3$, $(R-G)_5$ and $(R-G)_7$ are the genuine R-G values, but the R-G values of $(R-G)_2$, $(R-G)_4$, $(R-G)_6$ and $(R-G)_8$ are the estimated ones by using the method introduced in Section 2.4. Because the estimated R-G values are less reliable than the genuine R-G values, they will have a smaller contribution in computing \hat{d}_{\perp}^{r-g} . This is the reason that w_{\perp}^{r-g} is multiplied by 0.8 in Equation 2-14. The missing **R-G** value at the G sample location (Figure 4) is estimated by

$$\hat{d}_{\perp}^{r-g} = \begin{cases} \frac{(R-G)_4 + (R-G)_2}{2}, & \text{if } \frac{e_{\perp}^{r-g}}{e_{-}^{r-g}} > R, \\ \frac{(R-G)_1 + (R-G)_3}{2}, & \text{if } \frac{e_{-}^{r-g}}{e_{\perp}^{r-g}} > R, \\ w_{\perp}^{r-g} * \frac{(R-G)_1 + (R-G)_3}{2} + w_{-}^{r-g} * \frac{(R-G)_4 + (R-G)_2}{2}, & \text{otherwise}. \end{cases} \tag{2-15}$$

where e_{\perp}^{r-g} and e_{-}^{r-g} are the edge strengths in the vertical and horizontal directions. If $e_{\perp}^{r-g}/e_{-}^{r-g} > R$, it means there is horizontal texture on the location G; if $e_{-}^{r-g}/e_{\perp}^{r-g} > R$, it means there is vertical texture on the location G. Since the derivation of the missing vertical-class and the horizontal-class pixels are similar, we just introduce the interpolation method to estimate the missing vertical-class pixels here.

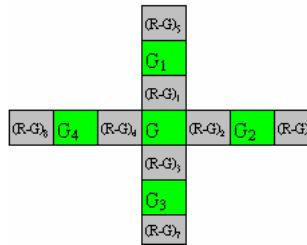


Fig. 4. The related pixels for estimating the R-G value of a missing vertical-class pixel at location G

After the interpolation operations mentioned above have been processed, the full **R-G** and **B-G** planes can be constructed. Accordingly, we can produce the demosaicked red plane (\bar{R}) from the **R-G** plane because $\bar{R} = \mathbf{R-G} + \bar{G}$. Similarly, the demosaicked blue plane (\bar{B}) can also be produced in the same way. Therefore, a full-color demosaicked image can eventually be constructed from the \bar{R} , \bar{G} and \bar{B} planes.

3 Experimental Results

In these experiments, we compared 24 test images, as shown in Figure 5, which are contained in the Kodak 3CCD image database. Each test image is a 24-bit full-color image containing one 8-bit value for each color and consisting of 512×768 pixels.



Fig. 5. The test images (from left to right, top to down, is image No.1 to No.24, respectively)

The test images were produced by preserving only one color component and abandoning the other two color components for each pixel from the original full color images. All the test images are represented as the Bayer CFA pattern. In our experiments, all test images were interpolated into demosaicked full-color images and their simulation results were compared by both the mean square error (MSE) and the peak signal-to-noise ratio (PSNR). MSE and PSNR are formulated as

$$MSE = \frac{|I_o - I_d|^2}{3 \times H \times W} , \quad PSNR = 10 \times \log_{10} \left(\frac{255^2}{(MSE)} \right) \quad (3-1)$$

where I_o and I_d are the original full-color image and its demosaicked full-color image, respectively; H and W are the height and the width of the original full-color image, respectively. For a demosaicked image, high fidelity implies high PSNR and small MSE measurements. Table 2 listed the measurement results of MSE and PSNR among ECDB, BI, ACPI, and ECI by using all of the 24 test images.

Table 2. Performance comparison of different methods

Image No	BI		ACPI		ECI		Proposed		ECDB	
	PSNR	MSE	PSNR	MSE	PSNR	MSE	PSNR	MSE	PSNR	MSE
1	26.230	154.910	33.780	27.232	33.816	27.007	35.844	17.159		
2	33.090	31.921	38.850	8.474	39.120	7.963	40.286	6.916		
3	34.360	23.828	40.660	5.586	40.211	6.194	41.738	4.493		
4	33.660	27.995	39.160	7.890	39.147	7.913	40.754	6.631		
5	26.720	138.382	35.040	20.374	35.044	20.355	36.708	14.073		
6	27.730	109.668	35.110	20.048	34.656	22.259	36.978	13.267		
7	33.470	29.247	40.850	5.347	40.448	5.866	41.863	4.415		
8	23.640	281.242	32.340	37.939	30.274	61.053	33.975	26.588		
9	32.510	36.482	40.330	6.027	39.427	7.420	41.581	4.617		
10	32.320	38.114	40.020	6.473	39.855	6.723	41.319	4.962		
11	29.300	76.398	36.340	15.104	36.336	15.118	38.270	9.794		
12	32.900	33.349	40.720	5.509	39.687	6.988	41.825	4.402		
13	23.950	261.867	29.930	66.082	31.034	51.248	32.097	40.271		
14	29.280	76.750	35.810	17.064	35.102	20.087	36.836	14.013		
15	31.500	46.034	37.440	11.724	38.200	9.842	38.734	9.152		
16	31.370	47.433	38.690	8.792	38.243	9.746	40.921	5.342		
17	31.960	41.408	38.670	8.832	39.177	7.859	40.183	6.272		
18	27.920	104.974	33.740	27.484	34.934	20.877	35.320	19.437		
19	28.160	99.330	37.310	12.080	35.046	20.346	39.062	8.228		
20	30.360	59.852	38.220	9.797	38.204	9.833	39.842	7.046		
21	28.590	89.966	35.350	18.971	35.587	17.965	37.285	12.356		
22	30.540	57.422	36.520	14.490	36.214	15.550	37.533	11.968		
23	35.300	19.190	42.050	4.056	41.478	4.627	42.641	3.676		
24	26.640	140.955	32.020	40.839	32.674	35.133	33.302	31.426		
Avg.	30.062	64.103	36.828	13.901	36.830	13.493	38.537	11.938		

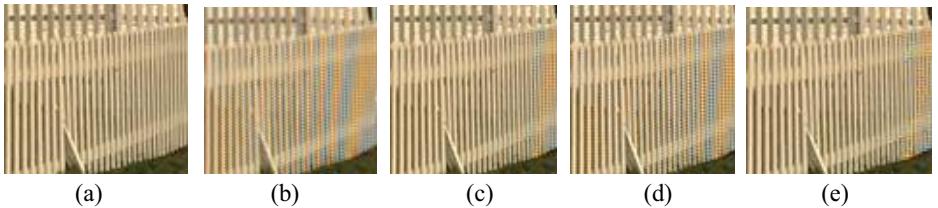


Fig. 6. Enlarged parts of the demosaicked image corresponding to the fence picture: (a) the original image, (b) BI, (c) ACPI, (d) ECI, (e) the proposed ECDB

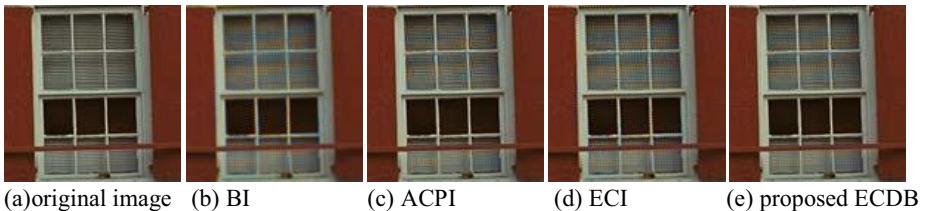


Fig. 7. Enlarged parts of the demosaicked image corresponding to the brickwood: (a) the original image, (b) GBI, (c) ACPI, (d) ECI, (e) the proposed ECDB

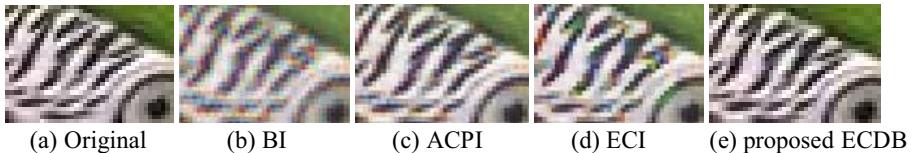


Fig. 8. Enlarged parts of the demosaicked image corresponding to the Parrots: (a) The original image, (b) BI, (c) ACPI, (d) ECI, (e) the proposed ECDB

Figures 6, 7 and 8 demonstrate the demosaicked results obtained by different demosaicking methods. In Figure 6 Figure 7 and Figure 8, both color artifact and zipper effect are mainly displayed in the fine-edge regions. Obviously, the ECDB algorithm provides the lowest color artifacts and the zipper effect for the edge region in the demosaicked result as shown in Figure 6 (e), Figure 7 (e) and Figure 8 (e). In the simulation results, the ECDB algorithm have the better performance and chosen suitable samples to interpolate missing pixels to reduce false colors, zipper effects, and then demosaicked images were more pleasing visual.

4 Conclusions and Future Work

In this paper, an effective CFA demosaicking interpolation algorithm is proposed which is called “Effective Color-Difference-Based Interpolation” (ECDB). The ECDB algorithm has provided pleasing results, and its performance is much better than the previous works shown in the experiments. Because the demosaicked image

produced by the ECDB algorithm still appears a few false colors at the high-contrast edge pixels, how to obtain more correct color at the high-contrast edge pixels has become the main research direction of our feature work.

References

- [1] Lukac, R., Plataniotis, K.N., Hatzinakos, D., Aleksic, M.: A new CFA interpolation framework. *Journal of Signal processing* 86, 1559–1579 (2006)
- [2] Tsai, C.-Y., Song, K.-T.: A new edge-adaptive demosaicing algorithm for color filter arrays. *Journal of image and vision computing* 25, 1495–1508 (2007)
- [3] Lee, C.-F., Pai, P.-Y., Huang, W.-H., Chang, C.-C.: An Effective Demosaicing and Zooming Algorithm for CFA Images. In: IEEE computer society of International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 1150–1153 (2008)
- [4] Pei, S.C., Tam, I.K.: Effective color interpolation in CCD color filter arrays using signal correlation. *IEEE Transactions on Circuits and Systems for Video Technology* 13(6), 503–513 (2003)
- [5] Chung, K.-L., Yang, W.-J., Yan, W.-M., Wang, C.-C.: Demosaicing of Color Filter Array Captured Images Using Gradient Edge Detection Masks and Adaptive Heterogeneity-Projection. *IEEE Transactions On Image Processing* 17(12), 2356–2367 (2008)
- [6] Newlin, D.R., et al.: An Efficient Adaptive Filtering for CFA Demosaicking. *International Journal on Computer Science and Engineering* 02(04), 954–958 (2010)
- [7] Lukac, R., Plataniotis, K.N.: Color filter arrays: Design and performance analysis. *IEEE Transactions on Consumer Electronics* 51(4), 1260–1267 (2005)
- [8] Kodak Lossless True Color Image Suite, <http://r0k.us/graphics/kodak/>

Utility Max-Min Fair Rate Allocation for Multiuser Multimedia Communications

Qing Zhang, Guizhong Liu, and Fan Li

School of Electronic and Information Engineering, Xi'an Jiaotong University,

Xi'an 710049, P.R. China

zhangqing@mailst.xjtu.edu.cn, {liugz,lifan}@mail.xjtu.edu.cn

Abstract. In this paper, we study the rate allocation problem among multiuser multimedia communications. Two essential objectives, efficiency and fairness, are often considered. The efficiency concerns how to maximize the sum of video qualities over all users, and the fairness concerns the video quality differences among users. Generally, increasing the efficiency and keeping the fairness are inconsistent. Towards this problem, we design a utility function by taking both the video quality and the allocated rate into consideration. Then, we propose a utility max-min fair rate allocation scheme, which can achieve a good tradeoff between the efficiency and the fairness. The simulation results demonstrate the validity of the proposed scheme.

Keywords: Multiuser multimedia communication, rate allocation, rate-quality model, quality fairness, utility max-min fairness.

1 Introduction

With the explosive growth of the Internet and the rapid advance of compression techniques, delay-sensitive multimedia networking applications such as video conferencing, Video on Demand (VOD), or Internet Protocol TV (IPTV) get more and more popular. Therefore, it is common that many video users share the same network bandwidth, and how to fairly and efficiently allocate the rate among them becomes more and more important.

The simplest multiuser rate allocation method is equally assigning the available network bandwidth to each user. A major problem of this method is that it does not consider the variable bit-rate characteristics of the video sequences. One method to overcome this disadvantage is that the controller collects the characteristics of all the video sequences and optimizes a global objective function using conventional optimization methods such as Lagrangian or dynamic programming [6]. For example, a commonly adopted method for the rate controller is to maximize the sum of the PSNRs

$$\max_{R_i} \sum_{i=1}^N PSNR_i(R_i), \text{ s.t. } \sum_{i=1}^N R_i \leq R \quad (1)$$

where R is the available network bandwidth and $PSNR_i$ is the PSNR of the i th user. However, this method ignores the fairness issue, which may result in a large

quality difference among multiple users. Maximizing the sum of the weighted PSNRs, wherein each PSNR of a video sequence is assigned a weight w_i , is a way towards fairness. However, in the literature, the weights w_i 's are usually heuristically determined, e.g., w_i is uniformly set to be $1/N$ [8].

Recently, the fairness issue is considered in multiuser rate allocation in a different setting. In [5], the authors formulated the optimal channel assignment problem as a convex optimization problem, aiming at a max-min fairness [1][7] for the downlink application. In [3], the authors proposed a generalized proportional fairness based on the Nash bargaining solutions and coalitions. While these two fairness criterions are successfully employed in networking applications, they cannot be directly used in content-aware multimedia applications since they do not explicitly consider the characteristics of the video content and the resulting impact on video quality.

In this paper, we consider the rate allocation problem among multiuser multimedia communications. We design a utility function by taking both the video quality and the allocated rate into consideration. Then, we propose a utility max-min fair rate allocation scheme, which can achieve a good tradeoff between the efficiency and the fairness. The rest of this paper is organized as follows. Section 2 presents the utility max-min fairness. Section 3 describes the algorithm to achieve the utility max-min fair rate allocation. Simulation results are given in Section 4, and we draw the conclusion remarks in Section 5.

2 Utility Max-Min Fairness Description

We consider N users that are transmitting video content in real-time over a shared channel/link. Since the channel has a limited bandwidth, it may not be able to satisfy the bandwidth requirements for all users. Therefore, a controller is applied to take charge of allocating the channel bandwidth to users $1, 2, \dots, N$.

2.1 Video Quality-Rate Model

In video compression, due to the quantization process, there exists a tradeoff between the distortion (D), which determines the channel bandwidth or storage space required to transmit or store the coded data. Generally, high bit-rate leads to small distortion while low bit-rate causes large distortion. Several models have been published in the literature to characterize this distortion-rate tradeoff for different video coders, such as MPEG-2, MPEG-4, and H.264. However, these models are usually used for rate control and cannot be used to model the distortion of an entire video coder for a given rate. In [9], a three-parameter model is proposed to feature the input-output behavior of the video coder. For the convenience of mathematical derivation, a two-parameter model is employed in [2]. However, this model cannot describe the measured distortion-rate performance of a video for a given coder with sufficient accuracy.

We find that video distortion-rate characteristics can be modeled by a two-parameter model as follows:

$$D(R) = aR^{-b} \quad (2)$$

where a and b are two positive parameters determined by the characteristics of the video content [10]. Since peak signal-to-noise ratio (PSNR) is a measure more common than MSE in the video coding and communication community, we use PSNR to measure the video quality (Q), which is calculated by

$$Q = PSNR = 10 \log_{10} \frac{255^2}{D} \quad (3)$$

By substituting Eq. (2) into Eq. (3), we get the quality-rate (Q-R) function of user u_i

$$Q_i(R_i) = \alpha_i + \beta_i \ln R_i \quad (4)$$

where $\alpha_i = 10 \log_{10}(255^2/a_i)$ and $\beta_i = 10b_i/\ln 10$.

2.2 Utility Max-Min Fairness Definition

As in [2], we assume that each user has a lowest desired quality constraint (lowest rate constraint R_i^L) and a highest satisfied quality constraint (highest rate constraint R_i^H). We assume that the available network bandwidth at least guarantees each user for the lowest desired rate R_i^L . Note that if the available network bandwidth is high enough to satisfy all the users with the highest satisfied rate R_i^H , the rate allocation problem is trivial since the controller just needs to allocate R_i^H to each user u_i . The rate allocation problem becomes challenge in the case that the available network bandwidth is not able to satisfy all the users with R_i^H .

To give a mathematical definition of utility max-min fairness, we first define the notion of feasible rate allocation.

Definition 1. A *feasible rate allocation* is a vector $\mathbf{R} = (R_1, R_2, \dots, R_N)$ that assigns rate R_i to user u_i such that

$$R_i^L \leq R_i \leq R_i^H, \quad (5)$$

$$\sum_{i=1}^N R_i \leq R, \quad (6)$$

where R is the available network bandwidth.

Definition 2. Given a feasible rate allocation vector $\mathbf{R} = (R_1, R_2, \dots, R_N)$, we define a corresponding utility-ordered vector $\mathbf{R}_o = (R_{i_1}, R_{i_2}, \dots, R_{i_N})$ such that $U_{i_k}(R_{i_k}) \leq U_{i_{k+1}}(R_{i_{k+1}})$ for $k = 1, \dots, N - 1$, where U_i is the utility function of user i . Given any other feasible rate allocation vector \mathbf{R}' and its utility-ordered vector $\mathbf{R}'_o = (R'_{j_1}, R'_{j_2}, \dots, R'_{j_N})$, we say $\mathbf{R} >_u \mathbf{R}'$ if and only if there exists some m such that $U_{i_k}(R_{i_k}) = U_{j_k}(R'_{j_k})$, for $0 \leq k < m$ and $U_{i_m}(R_{i_m}) > U_{j_m}(R'_{j_m})$. Note that the ordering is done over the space of feasible rate allocation vectors. However, to compare two vectors, we must examine their utility-order counterparts.

Definition 3. A utility max-min fair allocation is a feasible rate allocation vector that is largest under the ordering defined by $>_u$.

Informally, a vector \mathbf{R} is utility max-min fair if it is feasible and its utility $U_i(R_i)$ cannot be increased while maintaining feasibility, without decreasing utility $U_j(R_j)$ which satisfies $U_j(R_j) \leq U_i(R_i)$.

2.3 Video User's Utility Function

So far, we haven't yet define the video user's utility function. A direct way is using quality function as utility function, and performing rate allocation according to quality max-min fairness. However, due to the potential huge difference in rate-quality characteristics for different video sequences, this scheme may result in a huge difference in allocated rates for different users, and substantially decrease the sum of qualities. By taking both the video quality and the allocated rate into consideration, we define the utility function of user i as follows

$$U_i(Q_i, R_i) = Q_i + \sigma \ln R_i, \quad (7)$$

where σ is a non-negative tradeoff multiplier and its value controls the balance between the efficiency and the fairness. Note that if we let $\sigma = 0$, the utility function and the quality function are identical, and utility max-min fairness reduces to quality max-min fairness. Therefore, quality max-min fairness is a special case of our utility max-min fairness.

By substituting Eq. (4) into Eq. (7), the utility function of user i becomes

$$U_i(R_i) = \alpha_i + (\beta_i + \sigma) \ln R_i. \quad (8)$$

3 Utility Max-Min Fair Rate Allocation

In this section, we describe an algorithm to implement the utility max-min fair rate allocation. The controller constructs a virtual user to aid computation, which is not really joining in the rate allocation. In each iteration, giving a rate, the virtual user can calculate its utility. Let each user's utility function equal to the virtual user's utility, a corresponding rate can be resolved. The rate should be in the range of $[R_i^L, R_i^H]$. Otherwise, the rate is set to be R_i^L or R_i^H . The iteration is repeated until the available bandwidth is totally assigned to all the users. The algorithm is implemented in three stages, which are described as follows.

Stage 1. Initialization: The controller constructs a virtual user VU , and its utility function is

$$U_{VU}(R_{VU}) = \alpha_{VU} + (\beta_{VU} + \sigma) \ln R_{VU}. \quad (9)$$

where $\alpha_{VU} = (1/N) \sum_{i=1}^N \alpha_i$, and $\beta_{VU} = (1/N) \sum_{i=1}^N \beta_i$. Given the available bandwidth R , the tradeoff multiplier σ and iterative index $t = 0$, initialize R_{VU}^0 with R/N .

Stage 2. Rate Calculation:

- 1) The controller calculates $U_{VU}(R_{VU}^t)$ with R_{VU}^t by using Eq.(9) and let $U^t = U_{VU}(R_{VU}^t)$.

2) For each user i , suppose R_i^{tmp} be a rate satisfying

$$U_i(R_i^{tmp}) = U^t \quad (10)$$

By substituting Eq. (8) into Eq. (10), we can resolve R_i^{tmp} as

$$R_i^{tmp} = \exp\left(\frac{U^t - \alpha_i}{\beta_i + \sigma}\right). \quad (11)$$

With the feasible individual rate constraint in Eq. (5), user i 's rate demand R_i^t is

$$R_i^t = \max\left\{R_i^L, \min\left[\exp\left(\frac{U^t - \alpha_i}{\beta_i + \sigma}\right), R_i^H\right]\right\}. \quad (12)$$

3) The controller sums up all the rate demands $R_{total}^t = \sum_{i=1}^N R_i^t$ and compares R_{total}^t with R :

If R_{total}^t is not equivalent to R , set

$$R_{VU}^{t+1} = R_{VU}^t \left(1 - \frac{R_{total}^t - R}{R}\right), \quad (13)$$

$t = t + 1$, and go to 1).

Else, set $L = t$, and conclude the stage.

Stage 3. Rate Allocation Decision: At this stage, the controller allocates the rate $R_i^* = R_i^L$ to user i , and the utility max-min fair rate allocation vector is $\mathbf{R}^* = (R_1^*, \dots, R_N^*)$. We would like to emphasize that with \mathbf{R}^* , the available bandwidth R is fully allocated by the controller to all the users.

Recall that the tradeoff multiplier σ plays a key role in balancing the efficiency and the fairness. Intuitively, to achieve a good tradeoff, the value of σ should be properly determined to reflect the rate-quality characteristics of all the concurrent video streams. From Eq. 8, we observe that for each user i , σ is directly added to his/her rate-quality model parameter β_i . Therefore, we guess that the optimal value of σ can be jointly determined by all the β_i , $i = 1, \dots, N$. In this paper, we propose a simple formulation to calculate σ as

$$\sigma = \frac{1}{N} \sum_{i=1}^N \beta_i. \quad (14)$$

4 Simulation Results

To evaluate the proposed utility max-min rate allocation scheme, we conduct simulations on five video sequences. They are: Foreman, Carphone, Coastguard, Silent and Mobile in QCIF format. Note that these video sequences include slow, medium or fast motion, as well as smooth or complex scene. We use state-of-the-art H.264 JM14.2 [4] video codec to encode the video sequences. By changing the quantization parameter (QP) or using the rate control feature, we are able to compress the video sequences at different bit-rates and achieve different quality requirements.

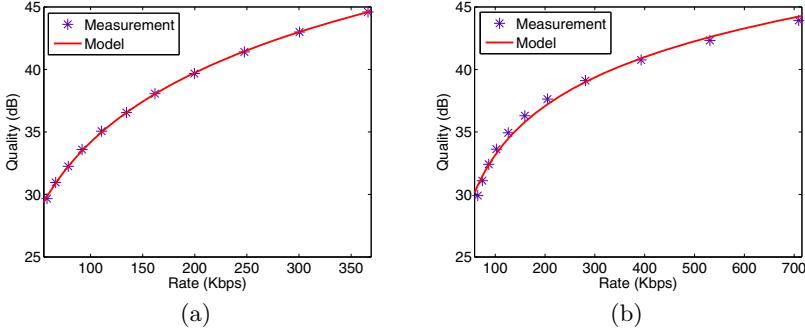


Fig. 1. Quality-rate curves for the test sequence: (a) Silent, (b) Foreman

Table 1. α_i, β_i, R_i^L and R_i^H for Different Video Sequences

Sequence	α_i	β_i	R_i^L (Kbps)	R_i^H (Kbps)
Foreman	7.1390	5.6500	57.1793	813.2653
Carphone	6.7610	6.2600	40.9478	449.6468
Coastguard	5.7910	5.2370	101.7666	1784.5382
Silent	-2.7590	8.0200	59.4218	385.6706
Mobile	-2.5110	6.0550	214.7078	2556.9433

4.1 Parameter Estimation

From Section 2, we can see that there are several parameters in our framework, α_i , β_i , R_i^L , and R_i^H . We can estimate α_i and β_i using offline training. For each video sequence, we first generate a set of (Q_i, R_i) by encoding the sequence using H.264 JM14.2 with different QP. Then, the optimal α_i and β_i can be computed by curve fitting tools, where the curve is featured by Eq. (4), and the data set needs to be fitted are the set of (Q_i, R_i) . We have shown the results of Silent and Foreman in Fig. 1. From the figure we can see that, with the optimal α_i and β_i , Eq. (4) can approximate quality-rate characteristics well. Similar results are observed for other sequences.

After finding the optimal α_i and β_i , we derive the values for R_i^L and R_i^H . As in [2], we suppose the lowest desired quality constraint Q^L is 30 dB, and the highest satisfied quality constraint Q^H is 45 dB. According to Eq. (4), we have

$$\begin{aligned} R_i^L &= \exp\left(\frac{Q^L - \alpha_i}{\beta_i}\right) \\ R_i^H &= \exp\left(\frac{Q^H - \alpha_i}{\beta_i}\right) \end{aligned} \quad (15)$$

The α_i , β_i , R_i^L and R_i^H for different sequences are shown in Table 1.

4.2 The Criterions of Efficiency and Fairness

Given a rate allocation vector $\mathbf{R} = (R_1, \dots, R_N)$, we evaluate it using the criterions of efficiency and fairness. The efficiency is calculated as the mean quality over all the users

$$\text{Efficiency} = \frac{1}{N} \sum_{i=1}^N Q_i, \quad (16)$$

where Q_i is calculated by Eq. (4). Clearly, the larger the mean quality, the higher the efficiency. We calculate the fairness as the standard deviation quality over all the users

$$\text{Fairness} = \sqrt{\frac{1}{N} \sum_{i=1}^N (Q_i - \bar{Q})^2}, \quad (17)$$

where $\bar{Q} = 1/N \sum_{i=1}^N Q_i$. Note that the standard deviation is no less than 0. If the qualities are absolutely fair, i.e., $Q_1 = Q_2 = \dots = Q_N$, then Fairness equals to 0, else Fairness is larger than 0. The more close to 0 the standard deviation, the better the fairness.

4.3 Multiuser Rate Allocation

We compare the proposed method with two approaches: the quality max-min fairness (QMMF), which maximizes the minimal quality of all the users, and the approach maximizing the sum of the PSNRs (MSPSNR), i.e., the traditional optimization-based approach shown in Eq. (II). Note that for QMMF and MSPSNR, the allocated rate should be within $[R_i^L, R_i^H]$. Otherwise, we set it to be R_i^L or R_i^H and re-allocate the rest rate for the other users. Given the video sequences to be transmitted, the available bandwidth R , the controller can figure out the rate allocated to each video sequence by using different methods, i.e., QMMF, MSPSNR, and the proposed method. Then, setting the allocated bit-rate as the target bit-rate, we compress the video sequence using the rate control feature in H.264 reference software JM14.2. Finally, each user transmits the compressed bitstream to the corresponding receiver. In the simulations, we consider five users, who transmit Foreman, Carphone, Coastguard, Silent, and Mobile to the corresponding receivers. Note that in the proposed method, unless explicitly specified, the tradeoff multiplier σ equals to 6.2444, which is calculated by Eq. (14) with the β_i s from Table II.

We test R at 1000, 1500, 2000, 2500 and 3000 Kbps. The allocation rate for each video sequence in different R using different methods are shown in Fig. 2. MSPSNR favors the video sequence that has a larger β since allocating more rate to the sequence with a larger β leads to a larger increase in the sum of the PSNRs. Specifically, the controller will first allocate each user with R_i^L . Then, the remaining rates will be first allocated to Silent until its R_i^H is reached. If there are some remaining rate, then Carphone will be satisfied first. QMMF tries to allocate more rate to the video sequence that has more complex motion and/or scene to guarantee that each user has the same video quality. Note that

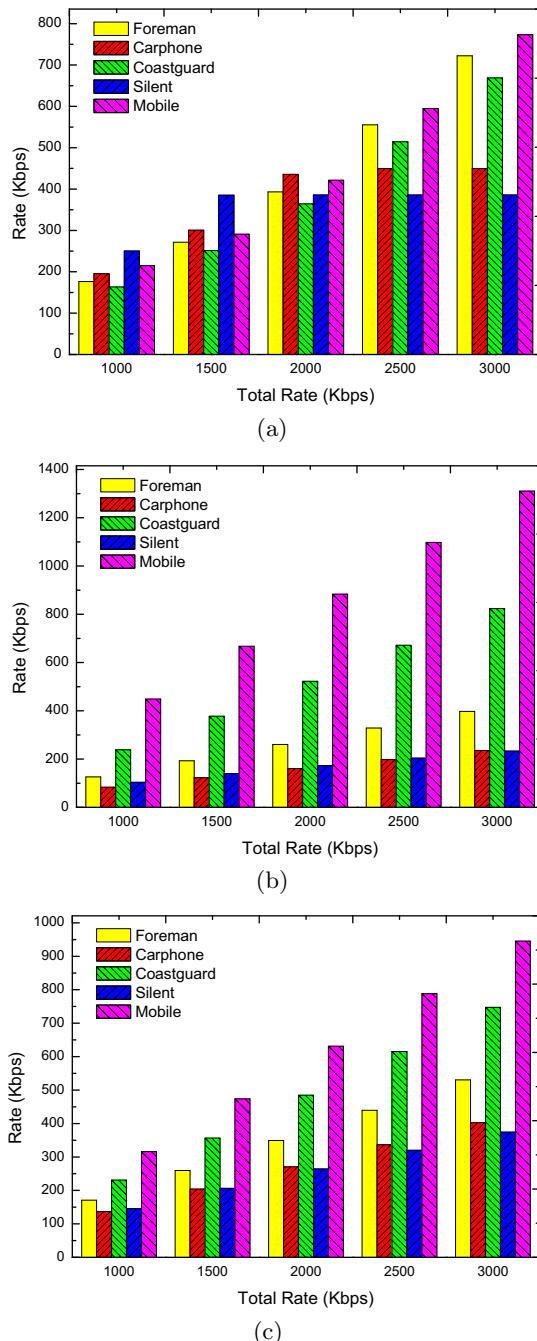


Fig. 2. Allocated rates for video sequences using different methods: (a) MSPSNR , (b) QMMF, (c) Proposed

the scene of Mobile is much more complex than that of Silent. To guarantee Mobile and Silent have the same quality, the controller has to allocate much more rate to Mobile than to Silent. By taking both the video quality and the allocated rate into account, the proposed method can balance the rate allocation among different video sequences. From Fig. 2 we can see that, at each R , for each sequence, the rate allocated by the proposed method is between the rate allocated by MSPSNR and the rate allocated by QMMF.

In Fig. 3(a), we show the Efficiency versus the available network bandwidth R . We can see that there is a big gap between the performance of QMMF and MSPSNR. The performance of the proposed method is inferior to that of MSPSNR, but is much better than that of QMMF. In Fig. 3(b), we show the Fairness versus the available network bandwidth R . The Fairness of QMMF is equivalent to 0 at each R , indicating that QMMF results in absolute fairness among all users. We also find that the Fairness of MSPSNR is fluctuant at different R , but

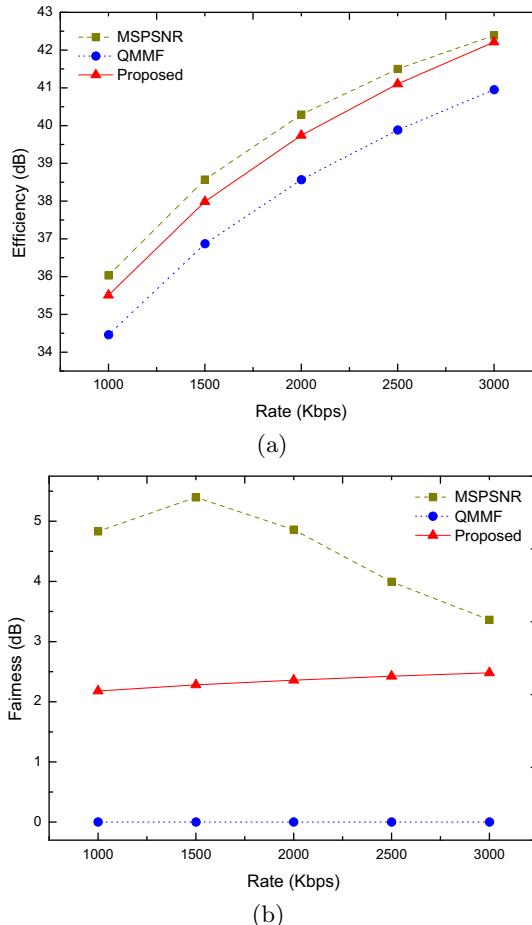


Fig. 3. Efficiency and Fairness of different methods: (a) Efficiency, (b) Fairness

Table 2. Efficiency and Fairness of the proposed method with different σ s

σ	1000 (Kbps)		2000 (Kbps)		3000 (Kbps)	
	Efficiency (dB)	Fairness (dB)	Efficiency (dB)	Fairness (dB)	Efficiency (dB)	Fairness (dB)
0.5	34.6507	0.3188	38.7744	0.3420	41.1733	0.3577
1	34.8055	0.5948	38.9454	0.6389	41.3550	0.6688
6.2444	35.5096	2.1780	39.7430	2.3597	42.2139	2.4814
10	35.6698	2.6904	39.9309	2.9229	42.3443	2.8876
100	35.9448	4.1150	40.2681	4.5415	42.3852	3.3454

generally at a high level. Recall that the more close to 0 the standard deviation, the better the fairness. In terms of fairness performance, although the proposed method is inferior to QMMF, it stably outperforms MSPSNR.

Finally, we investigate the impact of different σ on the performance of the proposed method. We test R at 1000, 2000 and 3000 Kbps, and the results are shown in Table 2. From the table we can see that, when σ is small, the proposed method has a good fairness and an unsatisfactory efficiency. Recall that QMMF, which is absolutely fair in terms of video quality, is a specialization of the proposed method with σ equaling to 0. As σ increases, the fairness gets worse while the efficiency increases. It should be mentioned that, the efficiency performance of the proposed method will be never superior to that of MSPSNR, even with a very large σ . This is because MSPSNR is a globally optimal rate allocation solution in terms of maximizing the efficiency. The proposed method aims at achieving a good tradeoff between the efficiency and the fairness, and this goal can be accomplished by choosing a proper σ in the proposed utility function. We find that in the illustration scenario, the performance of the proposed method with $\sigma = 6.2444$ is satisfactory under different available bandwidth, which is calculated by Eq. 14.

5 Conclusion

We propose a utility max-min fair rate allocation scheme for multiuser multimedia communications. The utility function is designed by jointly taking the video quality and the allocated rate into consideration. We propose an algorithm to accomplish the task of utility max-min fair rate allocation. We also explicitly formulate how to determine the tradeoff multiplier σ in the utility function. The simulation results show that with the calculated σ , a good tradeoff between the efficiency and the fairness can be achieved.

Acknowledgments

This work is supported in part by the National 973 Project No.2007CB311002, the National Natural Science Foundation of China (NSFC) Project No.60572045,

and the Ministry of Education of China Ph.D. Program Foundation Project No.20050698033.

References

1. Cao, Z., Zegura, E.: Utility max-min: An application-oriented bandwidth allocation scheme. In: Proc. IEEE INFOCOM (1999)
2. Chen, Y., Wang, B., Liu, K.J.R.: Multiuser rate allocation games for multimedia communications. *IEEE Trans. Multimedia* 11(6), 1170–1181 (2009)
3. Han, Z., Ji, Z., Liu, K.J.R.: Fair multiuser channel allocation for OFDMA networks using nash bargaining solutions and coalitions. *IEEE Trans. Commun.* 53(8), 1366–1376 (2005)
4. JVT Codec Reference Software, <http://iphom.hhi.de/suehring/tm1/download>
5. Mala, A., El-Kadi, M., Olariu, S., Todorova, P.: A fair resource allocation protocol for multimedia wireless networks. *IEEE Trans. Parallel Distrib. Syst.* 14(1), 63–71 (2003)
6. Ortega, A., Ramchandran, K.: Rate-distortion methods for image and video compression. *IEEE Signal Process. Mag.* 15(6), 25–50 (1998)
7. Radunovic, B., Le Boudec, J.: A unified framework for max-min and min-max fairness with applications. *IEEE/ACM Trans. Network.* 15(5), 1073–1083 (2007)
8. Shen, C., van der Schaar, M.: Optimal resource allocation for multimedia applications over multiaccess fading channels. *IEEE Trans. Wireless Commun.* 7(9), 3546–3557 (2008)
9. Stuhlmuller, K., Farber, N., Link, M., Girod, B.: Analysis of video transmission over lossy channels. *IEEE J. Sel. Areas Commun.* 18(6), 1012–1032 (2000)
10. Zhang, Q., Liu, G.: Rate allocation games in multiuser multimedia communications. *IET Communications* (Accepted for publication)

Adaptive Model for Robust Pedestrian Counting

Jingjing Liu, Jinqiao Wang, and Hanqing Lu

National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences
`{jjliu,jqwang,luhq}@nlpr.ia.ac.cn`

Abstract. Toward robust pedestrian counting with partly occlusion, we put forward a novel model-based approach for pedestrian detection. Our approach consists of two stages: pre-detection and verification. Firstly, based on a whole pedestrian model built up in advance, adaptive models are dynamically determined by the occlusion conditions of corresponding body parts. Thus, a heuristic approach with grid masks is proposed to examine visibility of certain body part. Using part models for template matching, we adopt an approximate branch structure for preliminary detection. Secondly, Bayesian framework is utilized to verify and optimize the pre-detection results. Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm is used to solve such problem of high dimensions. Experiments and comparison demonstrate promising application of the proposed approach.

Keywords: pedestrian counting, adaptive model, grid mask, RJMCMC.

1 Introduction

Pedestrian counting is sometimes required and essential in realms of multimedia and computer vision because of its wide-spread applications, such as video surveillance, driver assistance and scene analysis. However, it is a challenging task due to various clothing, body articulation and spatial occlusion and so on. For decades, research works and realistic applications on pedestrian counting have employed a myriad of methods, many of which rely on human detection. Hence it is indispensable to discuss the methods of pedestrian counting as well as the relevant ones for detection. Traditional methods pay more attention to static appearance of human, for instance texture, shape or silhouette. In the following, we will review these methods in detail.

Model-based methods build human models via shape or silhouette representation with prior knowledge. Lin *et al.* [1] proposed a hierarchical matching method based on artificially constructed models made up by part-templates to deal with occlusion. Zhao *et al.* [2] used 14 artificial 3D models to simulate specific poses of pedestrian, capturing side and front characters to handle body flexibility. Wu *et al.* [3] presented an approach which learned a series of certain part detectors for body parts through edgelet so as to improve detection rate.

The other methods train low-level feature based classifiers for human detection or counting whereas without premise of human body models. The low-level features reflect local details yet do not belong to any corporal concepts. Viola *et al.*

[\[4\]](#) learned a cascade structure classifier using textures of pedestrian appearance. Leibe *et al.* [\[5\]](#) trained visual codebooks to estimate spatial occurrence distribution of pedestrians. The well-known HOG descriptor proposed by Dalal *et al.* in [\[6\]](#) made use of local gradient information to depict appearance of human. A more recent method appears in [\[7\]](#). Gao *et al.* proposed a feature representation (ACF) similar to HOG descriptor and formed a classifier exploring the co-occurrence of discriminative features.

Besides the static appearance, motion information from video sequences is employed for pedestrian counting or relevant tasks. Here, relationships among sequential frames in video are all taken as motion information. Viola *et al.* [\[4\]](#) learned feature filters for final classifier, including filters based on different images of sequential images. Given camera is fixed, others used motion information for pre- or post- process, for example motion segmentation. Zhao *et al.* [\[2\]](#), [\[8\]](#) verified and optimized preliminary detection results with foreground mask under Bayesian framework. A.B. Chan *et al.* [\[9\]](#), [\[10\]](#) introduced Dynamic Texture (DT) into the field of crowd monitoring. As a pre-process, DT segmentation can not only capture the moving regions for the following features extraction and regression but also distinguish objects with different velocities.

Toward robust pedestrian counting with partly occlusion, we propose an adaptive model based approach for pre-detection. Three part models: head-shoulder (HS), right torso (RT) and left torso (LT) are established according to a pre-constructed pedestrian model. Assuming that there are only pedestrians standing or walking in the motion regions, head and shoulder are seldom occluded otherwise we do not take the sample as a valid pedestrian. Thus head-shoulder model is always contained in the adaptive model. In contrast, whether the model of left/right torso is reserved or discarded relies on visibility of the relevant body part in the image. Here, we propose a heuristic approach using low-level feature to determine the co-occurrence between head-shoulder and torso sides. With the adaptive model, we build up an approximate branch structure for integrate detection. In order to avoid penalty caused by deformation or articulation, we place template matching of torso sides before final classification using the adaptive model. Then, we verify and optimize the detection results under Bayesian framework on the basis of motion segmentation.

2 Adaptive Model

Model-based methods with prior knowledge try to link concept 'human'/'body part' to low-level features. Yet current methods with permanent part models usually have less flexibility thus bringing about improper punishment, in condition that some part detection get low scores due to occlusion or deformation. On the other hand, although some classifiers, for example the one learned by boosting algorithm, rely on training set and are sensitive to parameter changes and occlusion to some degree, low-level features based methods are expert in explaining local details. So it makes sense to take advantage of both model-based methods and low level feature based ones.

2.1 Part Models

In crowded scenes, pedestrians often occlude each other which adds difficulties for detection. Part detectors have been proved efficient in pedestrian detection meanwhile dealing with occlusion [1], [3]. In common sense, the head-shoulder, the left torso and right torso are seldom occluded simultaneously. Nevertheless, it is still unknown whether either torso side of pedestrian is occluded or which side is not. So it is not proper to construct a classifier of cascade structure [1] using part detectors or a classifier composed of part-templates with different weights [1]. Besides, it is not wisdom to ignore some persuasive evidence such as torso region which indicates existence of pedestrian.

The proposed adaptive model is defined as a set of part models corresponding to upper body parts of pedestrian. All the part models as well as legs model compose the whole pedestrian model. Dealing with occlusion, the adaptive model could tolerate some part models' missing but the others must be consisted. Therefore, distinct situations of occlusion are supposed to refer to different pedestrian adaptive models. As adaptive, the model tries to utilize as much convincing information as possible thus pursuing robust result. See Fig. 1(a), to detect such a person, adaptive model should only contain model of left torso (from perspective of picture viewers) while discarding the unconvincing information of right torso for it being vague. However, in Fig. 1(b), as being visible in scene, models of both torso sides should be adopted in final adaptive model.

Compared with color and texture, shape information is more confident to capture the characteristic of a pedestrian. However, body articulations and some equipment such backpack might bring about some mistakes or false alarm into the appearance. Thereby, we just consider the relatively invariant shape of human body with prior knowledge, neglecting the parts of high DOF (degree of freedom). A low dimensional shape model is constituted by 3 ellipses whose positions are relatively fixed: one indicates head, the others torso and legs. Fig. 2(a) illustrates integral shape model of a pedestrian with normal height, body proportion and fatness, similar to 2D model in [8]. We find that such a simple model is competent in practice.

Under the help of integral shape model, we separate the model into 3 regions (see Fig. 2(b)), which is similar to the division in [3]: head-shoulder, torso and legs. Our part models are different from [3], which include head-shoulder, left



Fig. 1. Different occlusion situations ought to correspond to different adaptive models.
(a) right torso part is vague; (b) both of the torso sides are visible in image.

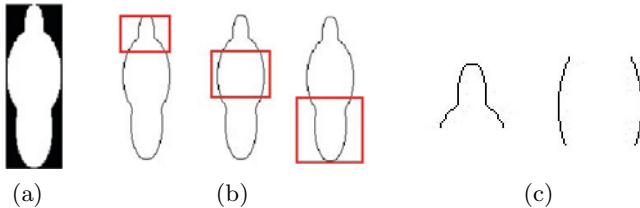


Fig. 2. Pedestrian model. (a) Full-body shape model; (b) regions of body parts: head-shoulder, torso and legs; (c) our part models: head-shoulder, left torso and right torso.

torso and right torso, as shown in Fig. 2(c). Each part model is represented by two parameter sets: positions of silhouette points and the unit image gradient vector set of the silhouette points. The introduce of gradient vector can partially handle problem of limited size change since unique model lacks scale variance.

2.2 Grid Mask for Torso Detection Using Consistent Contour

Because of size change, body articulations or occlusion, original part-template matching for torso detection is not robust. Consequently, it is crucial to avoid bringing improper punishment to final determination. To escape from this trap, we propose a heuristic approach based on contour consistence which searches existence of torso sides. Two grid-structure masks (as illustration in Fig. 3(a)) each containing 9 squares (in practice we use 3×3 or 5×5 size) are adopted to undertake this commission. For each mask, the red square marked by '00' is a start point. The mask with a top-right start detects left torso side whereas the other right side. Squares in green are the terminals. A path being found under defined consistent criterion between the start and the terminal square means consistence of torso sides edges in the squares' effective field.

The low-level features have shown good performance in explaining the local contour. So, we use a gradient feature similar to HOG descriptor [6] as the description of a square: gradient vectors of all points in the square's effective field merge into a unit vector, which represents the direction of the edge in the square. Since these features rely on a relative broad area but not pixels, this approach with grid masks is capable of tolerating with fat/thin, actions of upper body and contour deformation. See Fig. 3(b), all the possible paths covers a series of torso side edge. The algorithm through which we find a path in the masks is illustrated in Alg. II.

In practice, first of all, we use the head-shoulder model to get some candidate positions of head. Then for each candidate, we place the the grid masks at the torso regions relative to the head candidates. If a terminal is found in the mask, a torso side is supposed to be visual in the image and relevant part model should be contained in the adaptive model. See Fig. 2(a), left torso is detected while in Fig. 2(b) both torso sides are found.

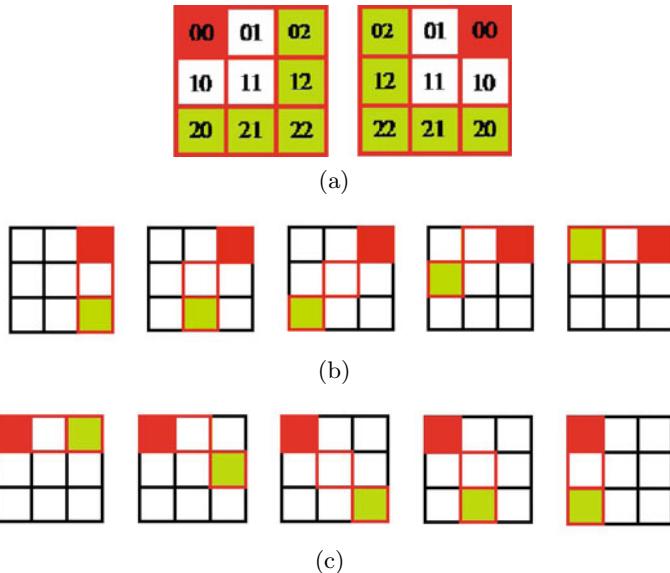


Fig. 3. (a) Grid masks for left and right detection; (b) consistent paths of left torso; (c) consistent paths of right paths

Algorithm 1. Torso side detection

1. Initialize the start coordinate and terminal coordinate:
 $(x_{start}, y_{start}) = (0, 0)$, $(x_{terminal}, y_{terminal}) = (0, 0)$
 angle between the unit gradient vector of start square and left/right shoulder vector of the constructed model D_{start} ought to be smaller than a threshold T_{torso}
 2. While $x_{terminal} \neq 2$ and $y_{terminal} \neq 2$
 - three coordinates of candidate squares :
 $(x_{start} + 1, y_{start})$, $(x_{start}, y_{start} + 1)$, $(x_{start} + 1, y_{start} + 1)$
 - select the **best candidate** square as the current end as well as the next mining start point
 $((x_{start}, y_{start}) = (x_{candidate}, y_{candidate}))$
 $(x_{terminal}, y_{terminal}) = (x_{candidate}, y_{candidate})$
 If there is no such square, the iteration stops.
 3. Either $x_{terminal}$ or $y_{terminal}$ if equal to 2, the left/right torso side is detected.
-

As a **best candidate** [7], we mean angle between unit gradient vector of the current start square and the best candidate square's direction defined above is the smallest among the three ones, besides lower than a threshold $T_{between}$. Besides, angle between the line connecting the two candidate's centers and the best candidate's direction ought to be larger than the sum $T_{consist}$.

3 Pedestrian Detection Based on Branch Structure

Inspired by the head candidate detection in [2], we adopt the part-template matching method along with contour gradients for body parts detection. Besides, motivated by efficiency of the cascade structure [11], we construct an approximate branch structure for pedestrian detection (see Fig. 4). Considering invariant shape and impossibility of occlusion of head-shoulder body part in this task, firstly, head-shoulder part-template matching is used to filter out lots of negatives. We set a relatively low threshold for this detection, pushing more candidates to pass through this access.

In the second step, the grid trigger decides which branch the candidates from the first step should go along. In practice, we find that even if the grid mask finds existence of a torso side, sometimes the score of torso side part-template matching is low because of deformation or poses of pedestrians, undermining the overall score of the adaptive model. This phenomenon obeys our expectation that is to employ warranted information as much as we could. To avoid such phenomenon, we place preliminary part-template matching of torso side on the branches before the adaptive model classification process. If some candidates could not get high score in the right/left torso template matching, the part model would not appear in the adaptive model. Although lose some information, these candidates could also possess a high possibility as a pedestrian because the only contained head-shoulder model might get a relative high score. Delicate thresholds achieve a good balance between grade of detect rate and decrease of false positive. Different branches result in different adaptive models.

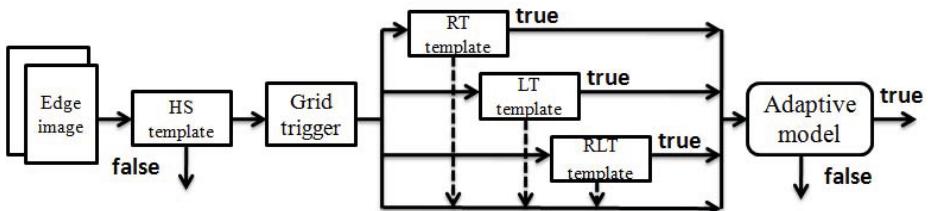


Fig. 4. The Approximate Branch Structure

Final classification in each branch relies on matching scores of part models contained in the adaptive model. Instead of fixed weights, we dispatch parts' weights according to their current scores . The higher is the score of the certain part, the bigger weight of this part, similar to making choices in realistic decision that human apt to accept information from more compromising channels.

4 Pedestrian Verification and Optimization

Since people sometimes share homogeneous velocities, it is hard to count pedestrians in crowded scene only with velocity. But motion information is no doubt

effective complement for people counting. Some methods have shown good performance in obtaining exact shape of motion regions. Additionally, we make use of motion information in Bayesian framework to verify and optimize the pre-detection instead of estimating overlapping rate.

4.1 The Bayesian Framework

Given a well segmented foreground image, an optimization problem is defined to find the object mask best covering the foreground image with the guide of the results from pre-detection. Achieve the solution equals to maximum a posterior (MAP) estimation:

$$O^* = \arg \max_{O \in \Theta} P(O|I) \quad (1)$$

$$P(O|I) \propto P(I|O)P(O) \quad (2)$$

where O denotes the objects mask with multiple models while I is the foreground image. Likelihood and prior are discussed in the following sections.

4.1.1 Prior Distribution

We use the integral shape of pedestrian model as mask of one object. In practice, an object has only two parameters: position and height (the ratio of height to width is fixed). In the object mask, objects model can be resized on the basis of the object's height. Hence, the prior probability of an object i contains the probability of its parameters and the punishing item of adding an object:

$$P(O) = \prod_{i=1}^n P(s_i)P(x_i, y_i)P(h_i) \quad (3)$$

$$P(s_i) = e^{\lambda_{punish} A_i} \quad (4)$$

The first probability relying on the object's size is a punishing item. A_i is area size of the object. λ_{punish} is set to handle different scenes of variant pedestrian density. More crowded the image, smaller the λ_{punish} . Suppose that pedestrian could appear at every position of the scene, $P(x_i, y_i)$ is uniform distribution while $P(h_i)$ is a Gaussian distribution $N(\mu_h, \sigma_h)$ whose expectation and variance refer to height of average people and real height fluctuation coming from statistics. Besides, the range of the height distribution h_{rang} depends on statistic parameters.

4.1.2 Likelihood

We accept the assumption that there are only pedestrians in the foreground scene and pixels in image are independent. Following the inference of [2], we easily reach likelihood of the MAP problem:

$$P(I|O) = \alpha e^{-(\lambda_{10}N_{10} + \lambda_{01}N_{01})} \quad (5)$$

where N_{10} is the number of pixels in foreground image which are not covered by the object mask. Comparatively, N_{01} is the number of pixels in object mask which covers background. Incremental computation [2] to compute N_{10} and N_{01}

is also adopted in our experiments. λ_{10} and λ_{01} are also set to compromise between two kinds of wrong overlapping mentioned above.

Using the prior Equ. 3 and the likelihood Equ. 5, the posterior probability is:

$$P(O|I) \propto \left(\prod_{i=1}^n e^{\lambda_{punish} A_i} e^{-\left(\frac{h_i - \mu_h}{\sigma_h}\right)} \cdot e^{-(\lambda_{10} N_{10} + \lambda_{01} N_{01})} ; h_i \in [1.5, 1.9] \quad (6)$$

4.2 RJMCMC for Pedestrian Counting

The solution of the MAP estimation includes the number of pedestrians, the positions and heights. There is no doubt that the solution space is of high dimensions. RJMCMC has demonstrated strong ability in searching for a solution in such a space [2]. We also utilize this algorithm to solve such MAP estimation. During iterations, a sampled candidate state in the solution space is accepted according to value of the Metropolis-Hastings acceptance ratio:

$$p(x, x') = \min\left(1, \frac{p(x')q(x', x)}{p(x)q(x, x')}\right) \quad (7)$$

where x and x' are the current and the candidate state, $p(\cdot)$ is the posterior distribution and $q(\cdot)$ is the proposal probability. Through iterations, the sampled states become dense and ideally the states converge to the solution. The number of crowd could be inferred from the solution. In pre-detection, the results offer probable pedestrian positions, which provide domain knowledge to the proposal probability. Actually, parameter (x_i, y_i) of states are sampled at these detections with little drifting, resulting in a fast convergence. Other parameters are sampled through iterations according to their distributions.

5 Experiments

Pedestrian counting experiments are carried on three video sequences from PETS 2009, which is a considerably challenging open dataset: S1.L1: 13-57-001 (seq.1), S2.L1: 12-34-001 (seq.2) and S2.L1: 12-34-008 (seq.3). Here, Seq.1 is a 221 frame sequence in which average number of pedestrian per frame is over 20, so people usually have severe occlusion. Seq. 2 contains 795 frames; average pedestrian number per frame is about 6. Seq.3 records the same scene as seq.2 but from a different viewpoint.

We take every pedestrian entering the scene as a valid pedestrian, until he leaves, even if he is totally occluded in the frame. In addition, we define what a good count is: we use position of head as a reference; the head position of a detection nearing the head top of a valid pedestrian with limited tolerance is seemed to be a true positive; however, if a correct count has been related to a pedestrian, other counts even if in the area of a good count, are seemed as false positives.

For motion segmentation, we employ Gaussian Mixture Models (GMMs) [12]. In every sequence, standard sizes of pedestrian model are re-estimated by size of height of a normal pedestrian in the scene and associated with μ_h . We multiply

Table 1. Compared results of performance evaluations

		Seq.1	Seq.2	Seq.3
Valid pedestrians		4719	4651	3856
Adaptive Model	Detection rate	65.75%	78.11%	67.66%
	False positive rate	5.11%	11.81%	17.17%
Approach in [2]	Detection rate	56.58%	79.54%	40.54%
	False positive rate	10.76%	11.08%	17.05%

**Fig. 5.** Some experiment results: from top to bottom related to seq.1, seq.2 and seq.3

a coefficient λ_{torso} to scores of torso sides template so as to compensate lose due to body articulation. In our experiments, crucial parameters are fixed as follow: $D_{start} = (0.707, 0.707)$, $T_{torso} = 0.9$, $T_{between} = 0.8$, $T_{consist} = 0.707$, $\lambda_{10} = 0.8$, $\lambda_{01} = 1.0$ and $\lambda_{torso} = 1.2$ (the three thresholds is recalculated by cosine operation according to angle). 1000 iteration of RJMCMC runs for each frame which means 4000 solution states are sampled. Our approach makes use of Bayesian framework with motion segmentation as [2] has done, however, we utilize the adaptive model for pre-detection, reaching a robust result toward pedestrian counting in crowd. Experimental results of our approach and comparison with

the approach in [2] are illustrated in Table. II. Similar work in [3] is also under Bayesian framework whereas ours is more concise for without training process.

Our approach represents an enhanced performance towards previous approach [2] in seq.1 and seq.3. Frames of seq.1 and seq.3 contain some occluded people and are selected to prove better ability of our approach in dealing with occlusion. Consider that the verification process can handle some occlusion and approach in [2] uses residue foreground analysis, the improvement on the database is precious. In seq.2, because of occlusion seldom happening, detection rates of the two approaches are close; ours even has a little bit decrease compared with approach in [2]. As discussed above, detected torso sides might get low score in template matching, causing improper punishment to total matching score, thus detection rate declines.

Some experiment results of our approach are illustrated in Fig. 5. Several factors prevent our method from better performance: (1) Ground truths are strictly labeled by people; (2) The strict constrain of a true positive leads to so many false positives. (3) The size of pedestrians usually flux so fiercely that a unique model varies in a limited scale range is not omnipotent. This phenomenon explains the high false positive rate. (4) GMMs used in our experiment just obtain the area which has apparent moving, the standing or slowly moving pedestrians are not reflected in the foreground mask.

For each 320×240 frame, our approach takes 0.2s with c++ code, satisfying a real-time application.

6 Conclusion

In this paper, a novel method named adaptive model have been proposed. Part models are not fixed or weighted depending on probability but on the inkling of body part existence. We put forward a heuristic algorithm with grid masks to detect torso sides using contour consistence. Then pre-detection are implemented with a classifier of branch structure. Detection results are than verified and optimized with RJMCMC. Experiments on an open database show promising results indicating our approach is robust in pedestrian counting in crowd scenes.

In the future, we are planning to make use of the prior knowledge in some method of feature training. So convert the original adaptive model to a new adaptive model, part models of which is semi-supervised learning.

Acknowledgement

This work is supported by National Natural Science Foundation of China (Grant No.60905008, 60833006) and National Basic Research Program (973) of China under contract No.2010CB327905.

References

1. Lin, Z., Davis, L.S., Doermann, D., DeMenthon, D.: Hierarchical part-template matching for human detection and segmentation. In: 11th IEEE International Conference on Computer Vision, Rio de Janeiro, Brazil, pp. 1–8 (2007)

2. Zhao, T., Nevatia, R.: Bayesian Human Segmentation in Crowded Situations. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA, pp. 459–466 (2003)
3. Wu, B., Nevatia, R.: Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detection. In: 11th IEEE International Conference on Computer Vision, Beijing, China, pp. 90–97 (2005)
4. Viola, P., Jones, M.J., Snow, D.: Detecting pedestrians using patterns of motion and appearance. In: 9th IEEE International Conference on Computer Vision, Nice, France, pp. 734–741 (2003)
5. Leibe, B., Seemann, E., Schiele, B.: Pedestrian detection in crowded scenes. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, USA, pp. 878–885 (2005)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, USA, pp. 454–461 (2005)
7. Gao, W., Ai, H.Z., Lao, S.H.: Adaptive contour features in oriented granular space for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, USA, pp. 1786–1793 (2009)
8. Zhao, T., Nevatia, R.: Stochastic human segmentation from a static camera. In: Workshop on Motion and Video Computing, Orlando, USA, pp. 9–14 (2002)
9. Chan, A.B., John Liang, Z.S., Vasconcelos, N.: Privacy preserving crowd monitoring: counting people without people models or tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Anchorage, USA, pp. 1–7 (2008)
10. Chan, A.B., Morrow, M., Vasconcelos, N.: Analysis of Crowd Scenes using Holistic Properties. In: 11th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Miami, USA (2009)
11. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai Marriott, USA, pp. 511–518 (2001)
12. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Ft. Collins, USA, pp. 246–252 (1999)
13. Ge, W., Collins, R.T.: Marked Point Processes for Crowd Counting. In: IEEE Conference on Computer Vision and Pattern Recognition, Miami, USA (2009)

Multi Objective Optimization Based Fast Motion Detector

Jia Su, Xin Wei, Xiaocong Jin, and Takeshi Ikenaga

Graduation school of Information, Production, System,
Waseda University, Japan
`selene@suou.waseda.jp,`
`weixin166@ruri.waseda.jp,`
`jxcking@moegi.waseda.jp,`
`ikenaga@waseda.jp`

Abstract. A large number of surveillance applications require fast action, and since many surveillance applications, motive objects contain most critical information. Fast detection algorithm system becomes a necessity. A problem in computer vision is the determination of weights for multiple objective function optimizations. In this paper we propose techniques for automatically determining the weights, and discuss their properties. The Min-Max Principle, which avoids the problems of extremely low or high weights, is introduced. Expressions are derived relating the optimal weights, objective function values, and total cost. Simulation results show, compared to the conventional work, it can achieve around 40% time saving and higher detection accuracy for both outdoor and indoor surveillance videos.

Keywords: Multi-objective optimization, Motion detection, Video surveillance.

1 Introduction

An important problem in computer vision is the determination of weights for multiple objective function optimizations. This problem arises naturally in many reconstruction problems, where one wishes to reconstruct a function belonging to a constrained class of signals based upon noisy observed data. There is usually a tradeoff between reconstructing a function that is true to the data, and one that is true to the constraints. Instances of this tradeoff can be found in shape from shading [1], Optical flow [2], surface interpolation [3], edge detection [4], visible surface reconstruction [5], and brightness-based stereo matching [6]. The recently popularized regularization method for solving ill-posed problems [7], [8] always requires the tradeoff of conflicting requirements. The basic framework defines a cost or error functional which reflects the "badness" of a proposed solution to the reconstruction problem. Mathematical techniques, such as the calculus of variations are used to find the best solution to the reconstruction problem. The contribution of each constraint to the cost functional is weighted, and the weights may be adjusted to achieve a desired tradeoff.

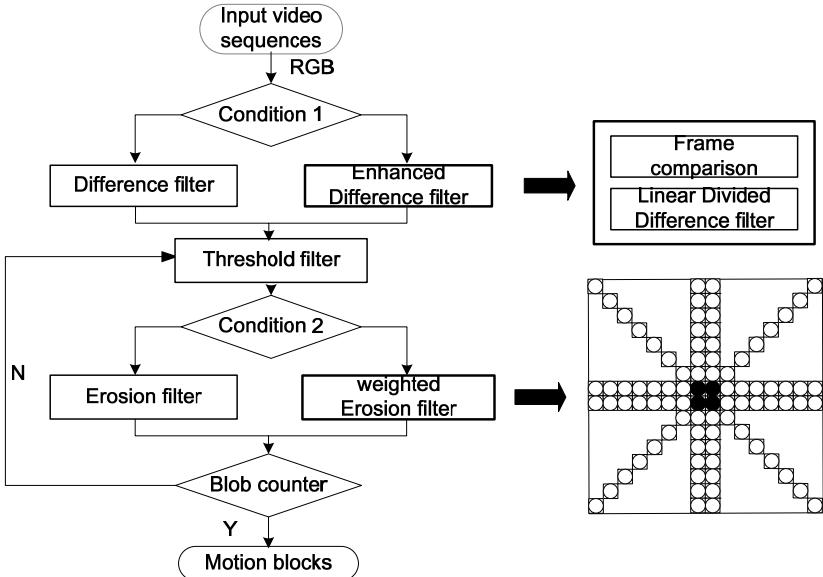


Fig. 1. Block diagram of the motion block detector

The motion detection embedded in the SAKBOT [9] system is based on background subtraction and models the background using statistics and knowledge-based assumptions. In fact, the background model is computed frame by frame by using a statistical function (temporal median) and taking into account the knowledge acquired about the scene in previous frames. In practice, the background model is updated differently if the considered pixel belongs to a previously detected moving visual object(MVO): in this case, the background model is kept unchanged because the current value is surely not describing the background. Moreover, if an object is detected as “stopped” (i.e., the tracking system detects that it was moving and is now stationary) for more than a “timeout” number of frames, its pixels are directly inserted into the background, without using the statistics.

Fig.1 displays the block diagram of the motion block detection. Different from other motion detection algorithm, it should be mentioned that the input video sequence is the raw data to be compressed. Video analysis algorithms can sustain a significant degradation in the input video quality without decrease in their accuracy. This finding leads to the problem of determining the tradeoff between video bit-rate and the accuracy of a given algorithm. Our approach to the solution of the problem is to focus on video features that affect performance of the algorithm. By studying how a video adaptation, which reduces video bit-rate, degrades video features, the rate--accuracy tradeoff can be estimated [10].

First of all, find the regions where these two frames are differing a bit. For this purpose, Difference and Threshold filters can be utilized. To remove random noisy pixels, the Erosion filter was chosen. The difference filter takes two images (source and overlay images) of the same size and pixel format and produces an image, where each pixel equals to absolute difference between corresponding pixels from provided images. The filter accepts 8 and 16 bpp(bit per pixel) gray-scale images and 24, 32,

48 and 64 bpp color images for processing. The threshold filter does image binarization using specified threshold value. All pixels with intensities equal or higher than threshold value are converted to white pixels. All other pixels with intensities below threshold value are converted to black pixels. The filter accepts 8 and 16 bpp grayscale images for processing. In order to simplify the model only gray level images were considered to get the motion blocks. Threshold is a technique that helps us to "delete" unwanted pixels from an image and only concentrate on the ones we want. For each pixel in an image if the pixel value is above a certain threshold convert it to 255 (white) otherwise converts it to 0 (black). Blob counter is a very useful feature and can be applied in many different applications. It can count objects on a binary image and extract them. The idea comes from "Connected components labeling," a filter that colors each separate object with a different color. Using BlobCounter we can get the number of objects, their position and the dimension on a binary image.

As the proposed idea utilizes both linear and nonlinear weight constraint in different motion detector parts, the paper has been written in 5 sections. First the principle of multi objective optimization has been illustrated; after that the optimized divided difference filter and weighted erosion filter have been described in detail. In the section 4 and 5 the simulation result and conclusion have been analyzed and drawn.

2 Multi Objective Optimization (MOO)

The Min-Max Principle First consider a simple problem. This problem will address the tradeoffs involved in a two-objective optimization problem, where a cost function is to be minimized over a single variable. Our goal is to shed some insight into the more complex general problem, where the cost function incorporate arbitrarily many objectives to be minimized over a multiple-variable field.

2.1 Linear Weight Constraint

Let $y_1(x)$ and $y_2(x)$ be non-negative functions of a single state variable x . It may be assumed, without loss of generality, that $y_1(x)=0$ and $y_2(x)=0$ for some (not necessarily the same) x . These are the two objective functions. Let the overall cost function be a linear combination of the objectives, defined by

$$e(\lambda, x) = \lambda y_1(x) + (1-\lambda) y_2(x), \quad 0 \leq \lambda \leq 1 \quad (1)$$

For a given value of λ , there will exist a value of x which minimizes $e(\lambda, x)$. Let that minimum value be $e^*(\lambda) = \min_x e(\lambda, x) = e(\lambda, x^*(\lambda))$. Given λ , one generally tries to $x^*(\lambda)$, the value with the minimum total cost. We can plot $e^*(\lambda)$ as a function of λ , since there exists some x such that $y_1(x)=0$, and the y_i 's are non-negative, $e^*(1) = \min_x y_1(x) = 0$. Likewise, $e^*(0) = 0$.

The total cost function must be convex, irrespective of the convexity of each objective function.

Let $e^*_1 = e^*(\lambda_1)$ and $e^*_2 = e^*(\lambda_2)$. Then for any α between 0 and 1,

$$e^*(\alpha\lambda_1 + (1-\alpha)\lambda_2) \geq \alpha e^*_1 + (1-\alpha)e^*_2. \quad (2)$$

The proof follows:

$$\begin{aligned} e^*(\alpha\lambda_1 + (1-\alpha)\lambda_2) &= \min_x e(\alpha\lambda_1 + (1-\alpha)\lambda_2, x) \\ &= \min_x e[\alpha(\lambda_1 y_1(x) + (1-\lambda_1)y_2(x)) + (1-\alpha)(\lambda_2 y_1(x) + (1-\lambda_2)y_2(x))] \\ &\geq \alpha \min_x [\lambda_1 y_1(x) + (1-\lambda_1)y_2(x)] + (1-\alpha) \min_x [\lambda_2 y_1(x) + (1-\lambda_2)y_2(x)] \\ &= \alpha e^*_1 + (1-\alpha)e^*_2 \end{aligned}$$

Note that the proof does not depend on the convexity of either y_i .

The main difficulty in choosing λ is that if it is either too low or too high, one of the objective functions will be inadequately represented in the total cost, and the total cost will be too low. One way to ensure that the total cost will not be too low is to pick the maximum cost solution. That is, find λ^* such that $e^* = e^*(\lambda^*)$ is maximized. This may seem at first to be far from optimal, but if one recalls that e has already been minimized over x , it will be seen that the maximization over x does make sense, while avoiding the problems of λ excessively low or high. This is the Min-Max Principle. Note that if there exists a value of λ not equal to zero or one for which the total cost is zero, then the optimal cost found using the Min-Max Principle will also be zero. This follows from the convexity of $e^*(\lambda)$.

What the extermination of the total cost implies is shown below. If λ^* is the weight which maximizes $e^*(\lambda)$, then denote the value of the corresponding state variable by $x^* = x^*(\lambda^*)$. Since $e^*(\lambda)$ is maximized at λ^* , conclude that

$$\begin{aligned} 0 &= \frac{d}{d\lambda} e^*(\lambda) \Big|_{\lambda=\lambda^*} \\ &= \frac{d}{d\lambda} e(\lambda, x^*(\lambda)) \Big|_{\lambda=\lambda^*} \\ &= \frac{\partial}{\partial \lambda} e(\lambda, x^*) \Big|_{\lambda=\lambda^*} + \frac{\partial}{\partial z} e(\lambda^*, x) \frac{d}{d\lambda} x^*(\lambda) \Big|_{\lambda=\lambda^*, x=x^*} \end{aligned}$$

But the last term, $\frac{\partial}{\partial \lambda} e(\lambda^*, x)$, equals zero, because x minimizes e . Therefore,

$$0 = \frac{\partial}{\partial \lambda} e(\lambda, x^*) \Big|_{\lambda=\lambda^*} = y_1(x^*) - y_2(x^*). \quad (3)$$

Each objective function assumes the same cost value, although the weights λ and $1-\lambda$ are not necessarily identical, and the weighted contribution of each objective function will not in general be identical.

2.2 Nonlinear Weight Constraint

In this section we consider a variation on the previous solution technique in which the total cost function is not a linear combination of the weights. Instead, let the weights of the objectives, when squared, sum to a constant. The constant can be set to 1 without loss of generality. That is,

$$e(\lambda, x) = \lambda y_1(x) + \sqrt{1-\lambda^2} y_2(x), \quad 0 \leq \lambda \leq 1 \quad (4)$$

As before, the optimal value of x given λ is $x^*(\lambda)$, where $e^*(\lambda) = \min_x e(\lambda, x) = e(\lambda, x^*(\lambda))$. A plot of $e^*(\lambda)$ would look much the same as in the linear combination case. It can be shown that $e^*(\lambda)$ is convex.

The Min-Max Principle requires that λ^* be found which maximizes the total cost, avoiding the problems of λ excessively low or high. As before, $e^*(\lambda) = \min_\lambda e^*(\lambda) = e^*(\lambda^*)$. Using the chain rule for differentiation we obtain:

$$0 = \frac{\partial}{\partial \lambda} e(\lambda, x^*) \Big|_{\lambda=\lambda^*} = y_1(x^*) - \frac{\lambda}{\sqrt{1-\lambda^2}} y_2(x^*). \quad (5)$$

After rearrangement,

$$y_1(x^*) = \frac{\lambda}{\sqrt{1-\lambda^2}} y_2(x^*). \quad (6)$$

Each objective function, when divided by its weight, is equal. The weights λ and $\sqrt{1-\lambda^2}$ will not in general be equal, therefore, the objective functions will not be equal. Also, the objective function with the greatest value will have the largest weight, so that its contribution to the total cost function is further increased. This contrasts with the linear weight case discussed above.

3 MOO Updated Divided Difference Filter

Compare the current frame with the first frame in the video sequence: if there were no objects in the initial frame, comparison of the current frame with the first one will lead to the whole moving object independently of its motion speed. The motion will be detected on the place where the car was. So the initial frame can be renewed sometimes, but still it will not give us good results in the cases where we cannot guarantee that the first frame will contain only static background. But, there can be an inverse situation, if a picture was put on the wall in the room. The motion detected until the initial frame should be renewed. Then there comes the condition 1.

Consider the general discrete-time nonlinear system [11]

$$\begin{aligned} x_{k+1} &= f(x_k, w_k) \\ y_k &= h(x_k, v_k) \end{aligned} \quad (7)$$

Where $x_k \in R^n$ is the $n \times 1$; state vector, $y_k \in R^m$ is the $m \times 1$ measurement vector, $w_k \in R^q$ the $q \times 1$; state noise process vector and $v_k \in R^r$ is the $r \times 1$ measurement

noise vector. It is assumed that the noise vectors are uncorrelated white Gaussian processed with expected means and covariance.

In this paper, because the measurement equation is linear, so the interval step-size is set: $h = 3$, (optimal for Gaussian distribution based on the square-root of the kurtosis). The state dimension is set as: $nx = \text{length}(x)$.

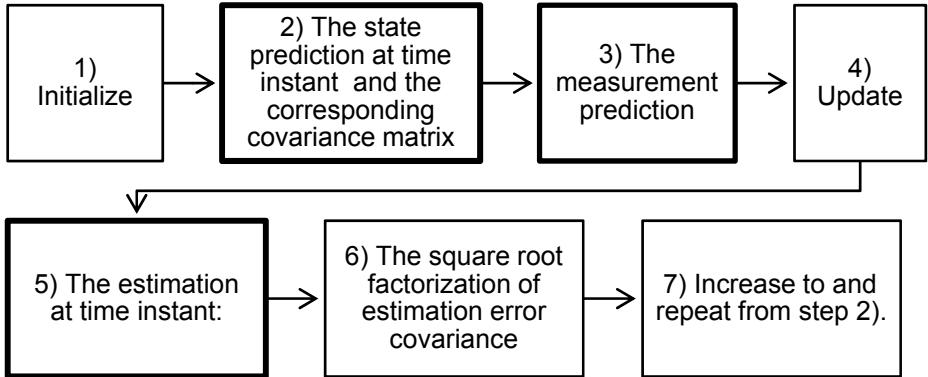


Fig. 2. Block diagram of the DDF

The main procedure of generate DDF is displayed in Fig. 2[12]

1) The state prediction at time instant $k+1$ and the corresponding covariance matrix is

$$\bar{x}_{k+1} = \frac{h^2 - n_x - n_w}{h^2} \psi(\hat{x}_k) + \frac{1}{2h^2} \sum_{p=1}^{n_x} \psi(\hat{x}_k + h\hat{S}_{x,p}) + \psi(\hat{x}_k - h\hat{S}_{x,p}) \quad (8)$$

2) The measurement prediction is

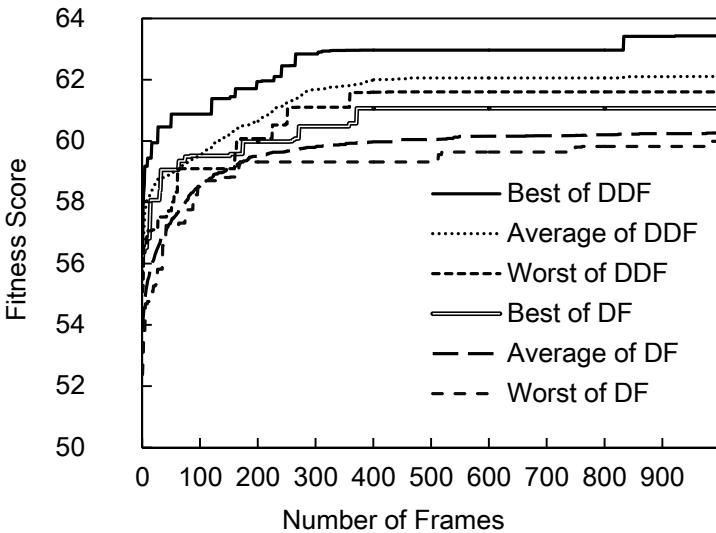
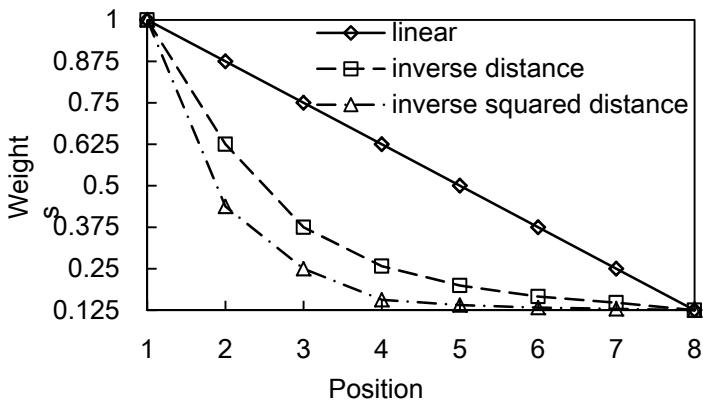
$$\hat{z}_{k+1/k} = H\hat{x}_{k+1/k} \quad (9)$$

3) The estimation at time instant $k+1$ given all the measurement up to time instant $k+1$ is:

$$\begin{aligned} \hat{x}_{k+1/k+1} &= \hat{x}_{k+1/k+1} + K_{k+1}(z_{k+1} - \hat{z}_{k+1/k}) \\ P_{k+1/k+1} &= (I - K_{k+1}H)P_{k+1/k+1} \end{aligned} \quad (10)$$

The noise function [13]

$$\text{abs}(\text{sum}(\text{noise}(2^i * x, 2^i * y, 2^i * z) / 2^i)) \quad (11)$$

**Fig. 3** Block diagram of the DF and DDF**Fig. 4.** Proposed adaptive erosion filter for a 16x16 macro-block

The noise clamp threshold parameter is useful to control the number of chips over the mesh: all the noise values smaller than the chosen threshold value are put to 0 (this means that also the displacement offset will be 0). The filter especially useful for binary image processing, where it removes pixels, which are not surrounded by specified amount of neighbors. It gives ability to remove noisy pixels (stand-alone pixels) or shrink objects. In Fig.3, DF means the conventional difference filter, DDF represents the divided difference filter.

Fig.4 displays the proposed adaptive erosion filter for a 16x16 macro-block, which contains two parts. The left paragraph shows the pixels of a macro-block. For example, the linear distance of the blackened pixels is 1. The right paragraph shows the different weights for the different distance between the picked pixels with the

central of each macro-block. The linear, inverse distance and the inverse squared distance which represent the x , y and z in noise function respectively. In order to collect every centralized noise, the weight should be decreased adaptively based on the different location of each pixel. By calculating the Blobs after the blob filter, the condition 2 has been set. The condition 2 is: if the $Num_{blobs} > Num_{objects}$, the weighted erosion filter is chosen; else the erosion filter is chosen. As the time discussion of different parts for this algorithm is illustrated in Table1, both condition 1 and 2 are considered both in accuracy and time consuming.

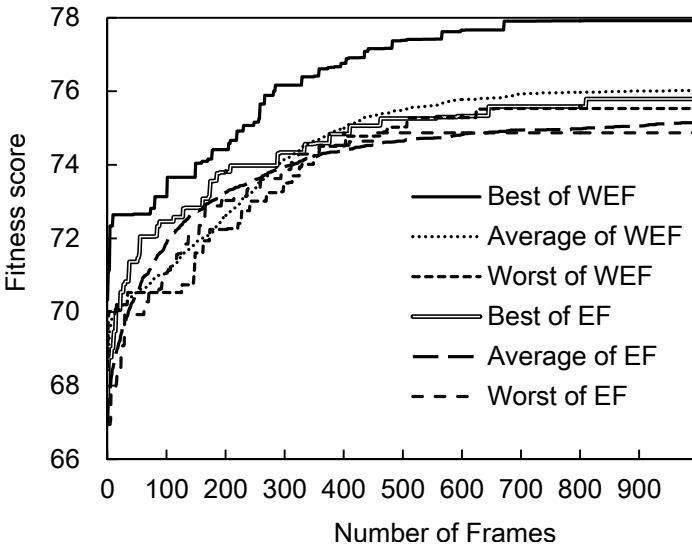


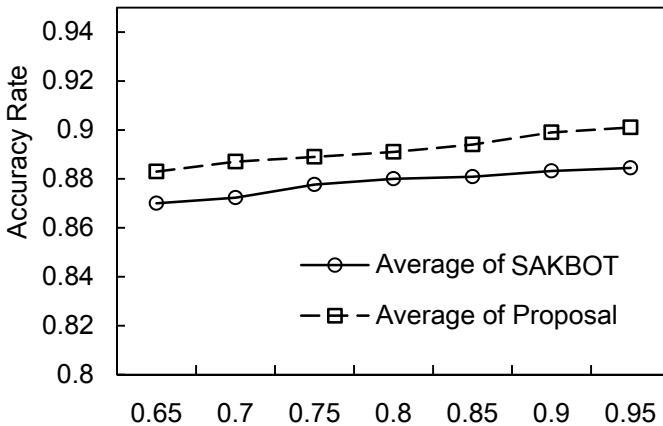
Fig. 5. Fitness Score of the EF and WEF

In Fig.5, EF means the conventional Erosion Filter, WEF represents the Weighted Erosion Filter.

4 Evaluation Result

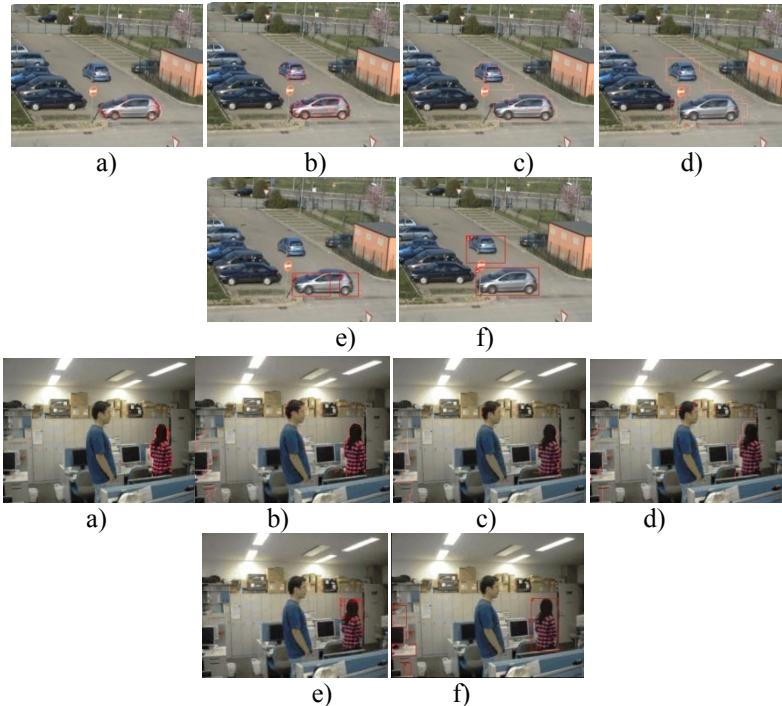
In order to evaluate the efficiency of the proposed motion block detector, the time comparison and the accuracy have been listed below in both Table 1 and Fig.6. Evaluation based on Ground Truth (GT) offers a framework for objective comparison of performance of alternate surveillance algorithm. This kind of evaluation techniques[14] compare the output of the algorithm with the GT obtained manually by drawing bounding boxes around objects, or marking up the pixel boundary of objects, or labeling objects of interest in original video stream. The standard CIF sequence (including 4 sequences for high noised, low noised, few motion, much motion) and visor[15] sequence for outdoor surveillance with 30 fps and a resolution of 352x576, to validate our proposed algorithm. Here, the T_{SAKBOT} and T_p means the motion block detector time of SAKBOT and proposed algorithm, respectively.

$$Time Saving = \frac{T_{SAKBOT} - T_p}{T_{SAKBOT}} \times 100\% \quad (12)$$

**Fig. 6.** Accuracy rate comparison with SAKBOT**Table 1.** Motion block detector time comparison

Sequences	SAKBOT(ms)	Proposal(ms)	Time Saving
visor_12054236	34359	19634	43%
49326_video00			
visor_12054236	24756	13528	45%
49326_video01			
visor_12054236	25995	15566	40%
49326_video02			
visor_12054236	30732	18078	41%
49326_video03			
visor_12066279	64031	38114	40%
10990_video_1			
visor_12066279	66309	37133	44%
10990_video_1			
visor_12066279	63042	34237	46%
10990_video_1			

Fig.7 shows the result of all the cases of updated filters. 8 outdoor and indoor surveillance sequences have been tested. Here, take two sequences as an example. From the comparison between the e) SAKBOT and f) Proposed Algorithm, the proposed algorithm can detected the moving cars and human accurately in both outdoor and indoor situation. Although from the indoor sequence, there cause some miss detection part on the left side of the sequence. The reason is mainly caused by the smoke detection module, which can also be respected by adding into proposed algorithm in the future work.



- a) Represents the detection result of DF+EF
- b) Represents the detection result of DDF+EF
- c) Represents the detection result of DF+WEF
- d) Represents the detection result of DDF+WEF
- e) Represents the detection result of SAKBOT
- f) Represents the detection result of Proposed Motion Detector

Fig. 7. Subjective Results comparing with SAKBOT

5 Conclusion

Adding the robust engine of motion block detector which combines a difference filter, erosion filter, divided difference filter and weighted erosion filter are proposed in this paper. The proposed method outperforms the original SAKBOT system counterpart in both time saving and motion detection accuracy performance in no smoking detection case. Theoretical analysis makes the divided difference filter into linear function which can easily embed in this motion detection engine. Namely, with the MOP technique, more lightening noise and uncritical parts were neglected, while more motion estimation computation could be saved. Moreover, the proposed approaches in this paper are easily to be applied in other kind of video analysis algorithm. Experimental results show that when these methods are integrated with proposed linear divided filter and weighted and erosion filter, around 2% of detection accuracy and 40% computation time can be saved with an acceptable coding quality loss.

Acknowledgments

This research was supported by “Ambient SoC Global COE Program of Waseda University” of the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

References

1. Horn, B.K.P.: Image Intensity Understanding. *Artificial Intelligence* 8(2), 201–231 (1977)
2. Horn, B.K.P., Schunck, B.G.: Determining Optical Flow. *Artificial Intelligence* 17(1-3), 185–203 (1981)
3. Grimson, W.E.L.: A Computational Theory of Visual Surface Interpolation. *Phil. Trans. Royal Soc. of London B* 298, 395–427 (1982)
4. Poggio, T., Voorhees, H.L., Yuille, A.: A Regularized Solution to Edge Detection. MITAI Laboratory Memo 833 (May 1985)
5. Cucchiara, R., Grana, C., Piccardi, M., Prati, A.: Detecting moving objects, ghosts and shadows in video streams. *IEEE Trans. Pattern Anal. Mach. Intell.* 25(10), 1337–1342 (2003)
6. Gennert, M.A.: Brightness-Based Stereo Matching. These Proceedings
7. Poggio, T., Torre, V.: Ill-Posed Problems and Regularization Analysis in Early Vision. MITAI Laboratory Memo773 (April 1984)
8. Terzopoulos, D.: Regularization of Inverse Problems Involving Discontinuities. *IEEE Trans. Pattern Analysis and Machine Intelligence* 8(4), 413–424 (1986)
9. Cucchiara, R., Grana, C., Prati, A., Vezzani, R.: Probabilistic Posture Classification for Human Behaviour Analysis. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans* 35(1), 42–54 (2005)
10. Su, J., Liu, Q., Ikenaga, T.: Motion Detection Based Motion Estimation Algorithm for Video Surveillance Application. In: International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2009) (2009)
11. Setoodeh, P., Khayatian, A., Farjah, E.: Attitude Estimation by Divided difference Filter-Based Sensor Fusion. *The journal of navigation* 60, 119–128 (2007)
12. Wu, C., Han, C.: Second-order Divided Difference Filter with Application to Ballistic Target Tracking. In: Proceedings of the 7th World Congress on Intelligent Control and Automation, June 25-27 (2008)
13. Su, J., Liu, Q., Ikenaga, T.: Lowbit-rate Motion block detection for uncompressed indoor surveillance. In: International Conference on Computational Science and Applications (ICCSA 2010) (March 2010)
14. Black, J., Ellis, T., Rosin, P.: A novel Method For Video Tracking Performance Evaluation. In: International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 125–132 (2003)
15. The ViSOR repository, <http://www.opensvisor.org>

Narrative Generation by Repurposing Digital Videos

Nick C. Tang¹, Hsiao-Rong Tyan², Chiou-Ting Hsu³, and Hong-Yuan Mark Liao¹

¹ Institute of Information Science, Academia Sinica

{nickctang, liao}@iis.sinica.edu.tw

² Dep. of Information and Computer Science, Chung Yuan Christian University

tyan@ice.cycu.edu.tw

³ Dep. of Computer Science, National Tsing Hua University

cthhsu@cs.nthu.edu.tw

Abstract. Storytelling and narrative creation are very popular research issues in the field of interactive media design. In this paper, we propose a framework for generating video narrative from existing videos which user only needs to involve in two steps: (1) select background video and avatars; (2) set up the movement and trajectory of avatars. To generate a realistic video narrative, several important steps have to be implemented. First, a video scene generation process is designed to generate a video mosaic. This video mosaic can be used as a basis for narrative planning. Second, an avatar preprocessing procedure with moderate avatar control technologies is designed to regulate a number of specific properties, such as the size or the length of constituent motion clips, and control the motion of avatars. Third, a layer merging algorithm and a spatiotemporal replacement algorithm are developed to ensure the visual quality of a generated video narrative. To demonstrate the efficacy of the proposed method, we generated several realistic video narratives from some chosen video clips and the results turned out to be visually pleasing.

Keywords: video narrative, video scene generation, video layering, motion interpolation/extrapolation.

1 Introduction

With the technology advance and cost down in hardware devices, more and more digital videos are available today and they have become important elements in our daily life. The explosive growth of digital videos shared on the websites such as YouTube attest people to express themselves visually. In addition, the improvement of video processing tools also motivates people to demonstrate their creativity by either modifying the visual content of digital videos, or merging video clips and etc. Consequently, storytelling and narrative creation have attracted attention from the research community. In [7], Lalonde *et al.* proposed a new framework to help user insert objects into an image to create a scenario. In [2], Chen *et al.* proposed a novel system in which a user only needs to perform freehand sketch and a set of photos can be automatically composed. However, to properly generate a narrative by composing atomic digital videos, both the seamless image blending and the spatiotemporal continuity issue have to be simultaneously considered. In this paper, we propose a new framework that can assist people in designing narratives just like telling a story.

To produce a video narrative, a number of challenging issues have to be handled. The first thing is video scene generation. For building a video scene, a two-step motion map construction algorithm is designed to estimate the motion flow of each block in every frame. The computed motion maps are not only used to maintain the temporal continuity in generated video narratives but also used as guidance to find out appropriate information to replace the area belongs to undesired foreground object. The second issue involved in a video narrative generation process is avatar preprocessing. The objects/avatars adopted for constructing a video narrative can be extracted from different video clips. For inserting the chosen avatars at different depths in a scene, their sizes needed to be regularized. Furthermore, we adopt motion interpolation and extrapolation algorithms to control the speed and movement of avatars and regularize the duration of avatars for generating a meaningful video narrative. Finally, for combining all layers to generate a realistic video narrative, a critical step that includes spatiotemporal placement and layer merging algorithm should be adopted. The spatiotemporal placement algorithm is used to place avatars in the narrative precisely. The layer merging algorithm is used to regulate the color and luminance properties of avatars to insure that combining all layers can be presented realistically.

The remainder of this paper is organized as follows. Section 2 describes the procedure of video scene generation. In Section 3, we present the proposed motion clip preprocessing and video narrative generation algorithm. Section 4 details the experiment results. Concluding remarks are drawn in section 5.

2 Video Scene Generation

The first step for generating a video narrative is to produce a visually pleasing video scene. Given a selected video clip which will be used to generate a video scene, we need to remove those undesired foreground objects first. Therefore we apply an efficient multi-resolution template matching technique to identify the undesired foreground objects in each video frame. Once the undesired foreground objects are tracked, we then compute the motion information of the remaining background areas to construct motion maps of each frame. Next, we perform patch referencing to locate appropriate background information from other frames to replace the areas of undesired foreground object through the assistance of motion vectors. In the meanwhile, a panoramic scene can be generated by stacking all the patches on spatial domain with the assistance of motion information on motion maps. In what follows, we shall report each step in details.

2.1 Construction of Motion Map

To maintain temporal continuity and to guarantee smooth video scene generation, we try to keep a motion map for every video frame. The procedure of computing a motion map is as follows. First, we apply a global motion estimation (GME) process to estimate the global motion between two adjacent frames. The result is then used by the local motion estimation (LME) process to reduce the computation complexity and improve the estimation accuracy.

Global Motion Estimation

Since SIFT can be accurately and efficiently applied in image matching, we use this technique to estimate the global motion between two consecutive video frames.

However, because both camera motion and object motion do contribute “motion” to videos, we need to incorporate a motion filtering mechanism into our GME algorithm. This is to ensure the mismatch of correspondences and the motion of moving objects do not influence the estimation results. The procedure of our proposed GME algorithm is as follows. First, we apply SIFT [9] to detect those salient match points between two consecutive frames and then introduce the RANSAC algorithm [6] to filter out the outliers which are caused by mismatches. The distance between each pair of remaining match points of two consecutive frames can be computed as candidate motions $V_{candidate}^k$. The definition of $V_{candidate}^k$ is as follows:

$$V_{candidate}^k = (x_F^k - x_{F+1}^k, y_F^k - y_{F+1}^k), \quad (1)$$

where (x_F^k, y_F^k) and (x_{F+1}^k, y_{F+1}^k) are the k_{th} pair of matching point between frame F and $F+1$. Then, we adopt a motion filtering mechanism to determine out the intrinsic motions $V_{intrinsic}^k$ where $(V_{intrinsic}^k - \bar{V}_{candidate}) < 2 \times \sigma_{V_{candidate}}$. Finally, we can compute $\bar{V}_{intrinsic}$ to represent the global motion V_g and then used as guidance in our local motion estimation procedure.

Local Motion Estimation

In this step, a correlation-based motion estimation algorithm with a correction mechanism is proposed to compute block motion vectors and removes undesired motion simultaneously. We use a larger block (16*16) to compute a motion vector, and then refine the result on a 8x8 block. The procedure of the proposed LME algorithm is as follows.

1. Use a modified CDHS algorithm proposed in [3] with the initial motion computed in the GME step to compute the block motion vectors for frame F_t .
2. Generate a pseudo frame F_{t+1}' based on the motion vectors calculated in step 1.
3. Compare the differences between F_{t+1} and F_{t+1}' . The differences can be viewed as a map of poorly estimated blocks.
4. For each poorly estimated block derived in step 3, re-estimate the motion vectors by the CDHS algorithm with the initial motion obtained from the surrounding valid motion vectors.
5. Construct a motion map based on the final motion vectors calculated for each block.

The constructed motion map of a frame indicates the motion flow of every block in that frame. This information can be used as a guideline when seeking available information from other frames and to maintain temporal continuity.

2.2 Object Removal by Patch Referencing

After constructing the motion maps of each frame, we then execute a patch referencing procedure to search available background information from other frames to replace the areas of undesired foreground object. To fill these regions by background information, the patch referencing process is triggered to search appropriate patches from other frames according to the motion information associated with the corresponding motion maps. Assuming the area of undesired foreground object of frame f is denoted as Ω , the surrounding region of Ω is denoted as $\delta\Omega$. We use the color

information of $\delta\Omega$ to search the most similar patch from other frames and then paste it into Ω directly. Figure 1 shows object removal result after executing the patch referencing process.



Fig. 1. Results of object removal. (a) original video frame. (b) result of object removal.

2.3 Panoramic Scene Construction

A panoramic scene can be built after all frames in this scene are inpainted. Since the areas covered by two consecutive frames can be determined based on the motion information embedded in motion maps, all patches can thus be stacked in the spatial domain. In this step, we apply the multi-resolution spline technique proposed by Burt and Adelson [1] to compose all patches to form a panoramic scene P . During the frame blending process, we also record the location of each frame in the mosaic P . Hence, P can be decomposed into several frames according to the relative locations of the frames in P and then merge with avatars to generate the video narrative.



Fig. 2. Results of panoramic scene generation. Video scene generated from a video with (a) panning camera motion and (b) shaky camera motion.

3 Video Narrative Generation

In this section, we shall describe how to generate a video narrative semi-automatically. Given a video narrative, a user has to choose avatars and uses a video scene as the basis for timing and positioning. Before going into the details of the video narrative generation process, we need to make some basic definitions.

To produce a video narrative, video clips are used as basic units. A video clip can be viewed as a spatiotemporally continuous list of frames and each frame is composed of several regions. The accumulation of these regions across temporal axis can form video layers. If video layers can be manipulated properly, a video narrative can be produced well. We give a few definitions here before our approach is discussed. A *video narrative* can be described as a 4-tuples (i.e., Time-index, X-coordinate, Y-coordinate, Layer) as follows:

$$\sum_{i=1}^n \sum_{a=0}^{\max_x} \sum_{b=0}^{\max_y} \sum_{c=1}^k (T_i, X_a, Y_b, L_c) \quad (2)$$

where n is the number of frames and k is the number of layers selected by the users, and \max_x and \max_y limit the screen size. Conceptually, a video schema is

4-dimentional, with layers representing one dimension. Given a video frame f_i at time T_i , $f_i = \sum_{a=0}^{\max_x} \sum_{b=0}^{\max_y} \sum_{c=1}^k (X_a, Y_b, L_c)$ can be decomposed into k regions. Stacking regions along the temporal axis results in a layer $l_w = \sum_{i=1}^n \sum_{a=0}^{\max_x} \sum_{b=0}^{\max_y} (T_i, X_a, Y_b)$, which is located at a series of screen positions (X_a, Y_b) with layer index w . A video scene, is used as basis to plan video narratives. A user can insert an additional layer/object $l_v = \sum_{i=j}^{j+n'} \sum_{a=0}^{\max_x} \sum_{b=0}^{\max_y} (T_i, X_a, Y_b)$ to the corresponding video scene, at positions (X_a, Y_b) from time instances j to $j + n'$, where n' is the duration of the additional layer. In what follows, we shall report each step in details.

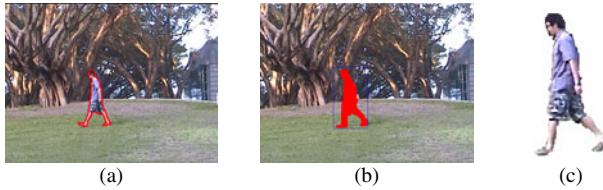


Fig. 3. Example of the avatar segmentation. (a) avatar selected in frame_t; (b) 1st round of object segmentation at frame_{t+1}; (c) after refinement.

3.1 Avatar Segmentation

Once a user picks an object to insert into the background scene, one of the most important tasks is to ensure that the object is extracted accurately. Besides, for generating a visually a pleasing video, a height estimation process is also introduced to estimate the size of an object more accurately. Next, we perform object segmentation to locate and extract foreground objects from each frame semi-automatically. The process starts with object selection, which is the only manual step in the whole process. As shown in Fig. 3(a), the object is selected manually in the first frame. Then, low-level attributes of the object are attached. These attributes include the center coordinates of object, (X_{oc}, Y_{oc}) , which will be used as the initial position in the search process, the color information, and the motion information. In this stage, we use the mean-shift color segmentation technique [4] to model the selected object, and use the segmented color information as input for the search process.

In the very beginning, the segmentation algorithm performs a full search of the neighborhood around the selected object centered at (X_{oc}, Y_{oc}) . After the object is tracked, we applied mean-shift color segmentation again on the targeted area and all blocks that contain the object's information are classified (as shown in Fig. 3(b)) based on the information collected in the object selection step. Then, we apply dilation and erosion on the contour of a segmented object and then use the results as inputs to an efficient matting algorithm [8] to obtain more precise results (as shown in Fig. 3(c)). Once an object is segmented accurately, we then estimate the height of an object based on [5] and use it as an input to regulate the size of the object at different depth levels.

3.2 Object Size Regulation

Viewing a video can be regarded as projecting 3D information in the real world onto a 2D screen. Hence, information will be lost during the projection process, such as the real size of an object, video depth, etc. to estimate the actual size of an object at different layers within a video scene and then use it to generate a realistic video narrative is an important issue. In other words, for positioning objects in motion clips to the video scene, we need to calculate the relative size of objects at different layers/depths. To solve this problem, we adjust the size of new avatars before insert them into the video scene.

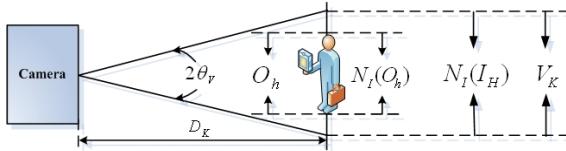


Fig. 4. Definition of video depth estimation

As illustrated in Fig. 4, the distance from camera to a measured surface (i.e., depth) and the actual height of object are denoted as D_k and O_h , respectively. The vertical angle of view is denoted as θ_v . The maximum value of measured actual height V_k can be expressed as follows:

$$V_k = \frac{N_I(I_H)}{N_I(O_h)} O_h \quad (3)$$

where O_h and I_H are the height of object O and image I , respectively. And, N_I is a function to compute the number of pixels in an image. According to Eq. (3), the depth D_k can be calculated by the following formula:

$$D_k = \frac{1}{2} V_k \tan \theta_v = \frac{1}{2} \left[\frac{N_I(I_H)}{N_I(O_h)} O_h \right] \tan \theta_v \quad (4)$$

Eq. (4) can be rewritten as:

$$N_I(O_h) = \frac{O_h N_I(I_H) \tan \theta_v}{2 D_k} \quad (5)$$

Since the resolution and the actual height of a selected object are fixed in a video clip, I_H and O_h can be considered as constant values in all depth layers. In addition, O_h can be determined in the object segmentation procedure. Therefore, we can regulate the size of an object based on the depth D_k . The limitation of our method is that the selected video should contain at least a foreground object. Figure 5 shows an example related to object size adjustment. Figure 5(a) shows that we insert an object into a video scene without size adjustment. Figure 5(b) shows the result after we regulate the size of the object.

3.3 Motion Interpolation and Extrapolation of Avatars

To generate a realistic video narrative, motion interpolation and motion extrapolation algorithms have to be introduced. In what follows, we shall describe how motion interpolation and motion extrapolation are implemented.

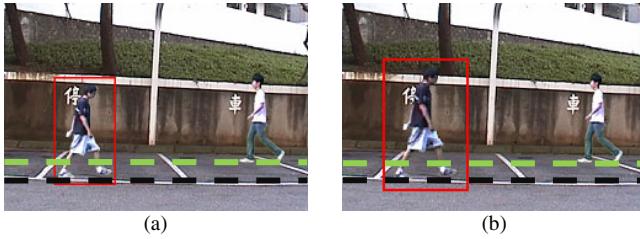


Fig. 5. Example of object size adjustment. (a) insert an object directly; (b) result of size adjustment.

Motion Interpolation

Usually, motion interpolation is applied whenever the number of motions is insufficient. This situation commonly happens in the case of changing the speed of avatars. For enriching the motions of an original motion clip, we applied our earlier work [12] to generate new poses of object and insert them into the motion clip. We briefly summarize this work as follows. First, we extract stick figures and contours of each posture in a motion clip. The stick figure is then used as guidance to analyze the motion of each part of a posture. A cycle prediction procedure based on computed contours is used to estimate the motion cycle r in a motion clip. Second, for generating a new posture $P_{i+0.5}$, we copy all available patches according to the computed motion flows from the neighboring postures P_{i+1} , P_{i-1} and P_{i+r} and then paste them into $P_{i+0.5}$. Finally, we apply a 2-D+t video inpainting technique and use patches inserted in previous step as guidance to inpaint the missing part of $P_{i+0.5}$.

Motion Extrapolation

Motion extrapolation is commonly used when a user adds an avatar into a narrative and he/she wants to control the movement of the avatar. Another situation that needs motion extrapolation is when the length of a motion clip is too short to generate a meaningful narrative. Under these circumstances, motion extrapolation can help extend a motion clip by duplicating motion cycles. In this work, we make use of a two-step algorithm which includes motion extension and object size normalization to generate new motion in our motion extrapolation procedure. The motion extension algorithm is described as follows.

1. The binary images of the last five motions in motion clip MC_1 and the first five motions in motion clip MC_2 are stacked along the temporal axis to produce motion stack MS_1 and MS_2 respectively.
2. Use MS_1 and MS_2 as inputs to find the positions, P_i^1 and P_i^2 , of the most similar motion in motion clip MC_i which is contained in our motion clip database.
3. Find the best-matched results in motion clip MC_i , and then use the motion flows between P_i^1 and P_i^2 to generate a new motion clip MC_e (as shown in Fig. 6(c)) which can be bridge the gap between MC_1 and MC_2 to make the new motion clip look continuous and homogeneous.

Since the extended motion clip MC_e is chosen, the size of the objects inside MC_1 , MC_2 and MC_e has to be normalized by using a bi-linear interpolation technique. After the adjustment of the object size, the two targeted video clips can then be combined.

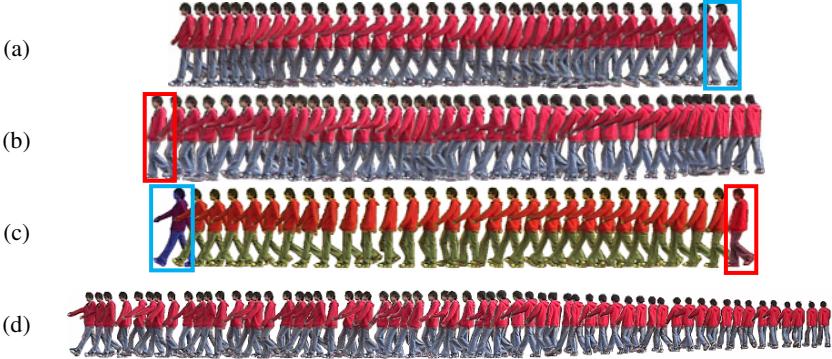


Fig. 6. An example of motion extrapolation. (a) Motion clip MC_1 ; (b) Motion clip MC_2 ; (c) Extended motion MC_e ; (d) an example of extrapolated motion clip.

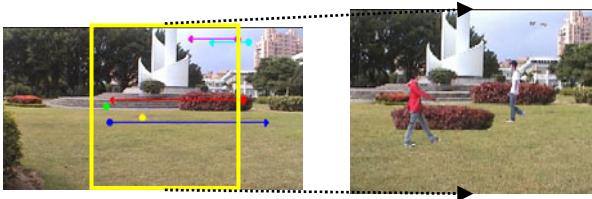


Fig. 7. Example of a designed narrative

3.4 Spatiotemporal Placement and Layer Merging

For generating a realistic video narrative, providing a spatiotemporal placement of motion clips and defining how motion tracks can be merged are considered. In our implementation, the spatiotemporal placement algorithm takes several steps. First, a user has to select a motion clip, decides where it starts, and then provides extra parameters such as a duration of time. Figure 7 shows an example of a designed narrative. Second, for each motion track, the time durations for showing up and the position of an object in a video frame need to be computed based on the settings defined in the first step. Computing the timing (i.e., frame number) for placing avatar can be accomplished by slicing the panorama. Besides, to make the avatars go smoothly, the camera motion of a video should be taken into consideration.

As shown in Fig. 8, assuming that F_0 and F_1 are two consecutive frames, where the target object (indicated by a smiling face in figure 8) is placed. Let (RTx, RTy) be a transformation matrix which integrates changes (including translation and rotation). Let (Tx, Ty) represent a translation vector of camera motion between F_0 and F_1 . The final location of the target object in F_1 is $(X'_o, Y'_o) = (X_o, Y_o) + (Tx, Ty)$. And the final location of the avatar is $(X_i, Y_i) = (X'_o, Y'_o) + (RTx, RTy) = (X_o, Y_o) + (Tx, Ty) + (RTx, RTy)$.

Recently, Poisson image editing [10] is considered as a popular and robust technique for seamless image composition. In [7], Lalonde et al. modified this concept to prevent severe discoloration and used a blending mask to insert objects into an image seamlessly. In [2], Chen et al. improved [10] with boundary condition and used alpha matte to produce better blending results. However, the results may not be always acceptable. For example, the target objects in a video may have excessive tone

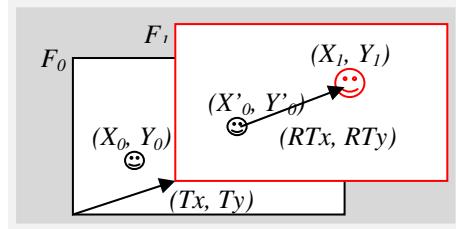


Fig. 8. Illustration of spatiotemporal placement



Fig. 9. Example of layer merging. Insert an avatar with (a) and without (b) tone adjustment algorithm.

changes. To solve this problem, we propose an algorithm to preserve object color information and make the result of layer merging more realistic. This algorithm can be separated into two steps. The first step is to use Poisson equation to blend the boundary part of objects within a target region; Then, in the second step, we adjust the color of each pixel in the object image based on the concept proposed in [11]. We assume that the area of an object to be inserted is denoted as O , the surrounding area of O is denoted as O_Ω and the target area in the video scene is denoted as O_Φ . We first apply Poisson image editing technique [10] to blend O_Ω with O_Φ . After that, we adjust the color of each pixel p in O_Ω by

$$C'(p) = (C(p) - M(O_\Phi)) \frac{D(O_\Phi)}{D(O_\Omega)} + M(O_\Omega) \quad (6)$$

where $M(O_\Phi)$ denotes the mean of pixel value and $D(O_\Phi)$ denotes the standard deviation of pixel value in the surrounding area. $M(O_\Omega)$ and $D(O_\Omega)$ are also defined for the inserted object. Therefore, the blending result of O can be presented by

$$O' = P(p) \cup C'(p) \quad (7)$$

where P is the function of Poisson Image Editing. In Fig. 9, we show the results of combining O_Φ and O_Ω using $C'(p)$ and $P(p)$ (Fig. 9(a)) and without using $C'(p)$ and $P(p)$ (Fig. 9(b)), respectively.

4 Experiment Results

To test the effectiveness of our new video narrative generation method, we conducted a series of experiments. In each experiment, we combined different layers from at

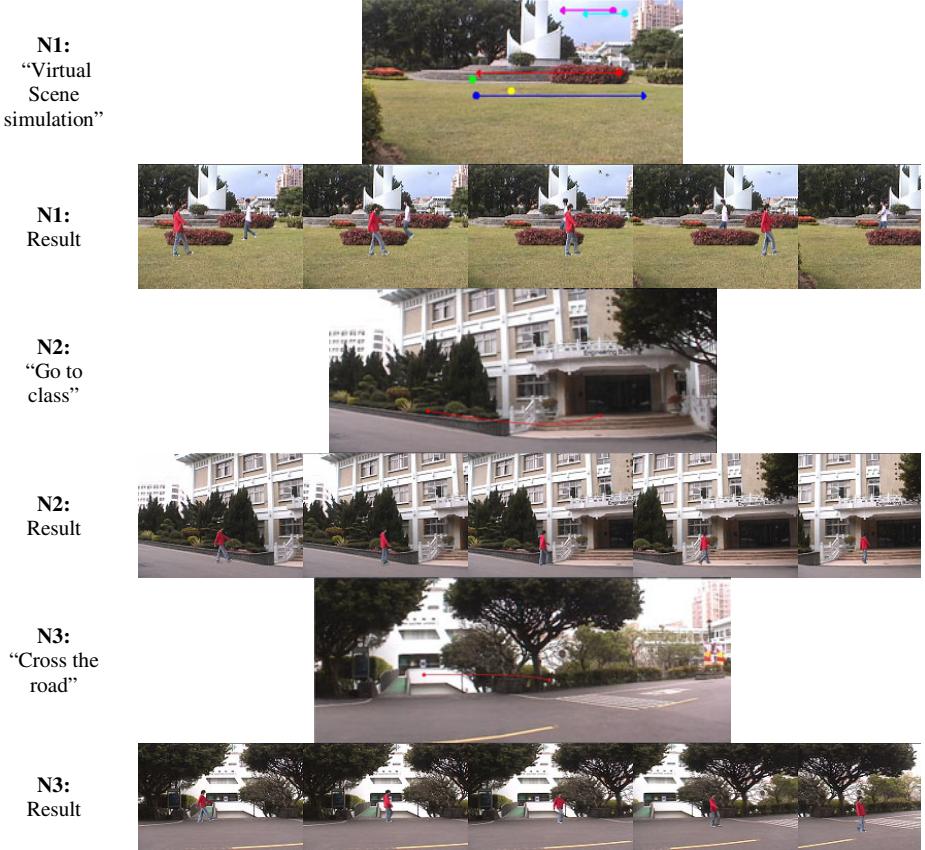


Fig. 10. Experiment results

least two video clips. In Fig. 10, N1 shows that several avatars such as bush, birds and human subjects are inserted into a natural scene successfully. N2 and N3 are two sets of video narratives which demonstrate two kinds of motion are combined, interpolated, and then inserted into different layers of video scenes. We demonstrate video narratives generated. Readers are recommended to visit our website at <http://research.twnct.net/VNG/>.

5 Conclusion

We have proposed a new framework which includes three major processes for generating video narratives from existing video clips. The first one is video scene generation. A video scene is used as basis to plan a video narrative. For producing a video scene, we first adopt a patch referencing algorithm to remove undesired objects and then stack all patches along the spatial axis to construct the video scene. Second, an

avatar preprocessing process is used to regulate a number of specific properties, such as the size of objects or the length of motion clips for producing a meaningful video narrative. Finally, an important procedure that includes a layer merging and a spatio-temporal replacement algorithm is adopted to ensure the visual quality of a realistic video narrative. The contribution of this technique is that users only need a few steps to design their own narratives. This is just like telling a story.

References

1. Burt, P.J., Adelson, A.H.: A Multiresolution Spline With Application to Image Mosaics. *ACM Trans. on Graphics* 2, 217–236 (1983)
2. Chen, T., Cheng, M.M., Tan, P., Shamir, A., Hu1, S.-M.: Sketch2Photo: Internet Image Montage. In: ACM SIGGRAPH ASIA (2009)
3. Cheung, C.-H., Po, L.-M.: Novel cross-diamond-hexagonal search algorithms for fast block motion estimation. *IEEE Trans. on Multimedia* 7(1), 16–22 (2005)
4. Comaniciu, D., Meer, P.: Mean Shift: A Robust Approach toward Feature Space Analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24(5) (May 2002)
5. Criminisi, A.: Single-View Metrology: Algorithms and Applications. In: Van Gool, L. (ed.) DAGM 2002. LNCS, vol. 2449, pp. 224–239. Springer, Heidelberg (2002)
6. Fischler, M., Bolles, R.: Random sample consensus: A paradigm for model fitting with application to image analysis and automated cartography. *Communications of the ACM* 24, 381–395 (1981)
7. Lalonde, J.F., Hoeim, D., Efros, A.A., Rother, C., Winn, J., Criminisi, A.: Photo Clip Art. *ACM Transactions on Graphics (SIGGRAPH 2007)* 26(3) (August 2007)
8. Levin, A., Lischinski, D., Weiss, Y.: A Closed Form Solution to Natural Image Matting. In: Int. Prof. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 61–68 (June 2007)
9. Lowe, D.G.: Object recognition from local scale-invariant features. In: Int. Conf. on Computer Vision, pp. 1150–1157 (September 1999)
10. Perez, P., Gangnet, M., Blake, A.: Poisson image editing. In: Proc. of ACM SIGGRAPH, pp. 313–318 (2003)
11. Reinhard, E., Adhikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. In: Proc. Conf. on IEEE Computer Graphics and Applications, vol. 21(5), pp. 34–41 (2001)
12. Shih, T.K., Tang, N.C., Tsai, J.C., Zhong, H.-Y.: Video Falsifying by Motion Interpolation and Inpainting. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2008)
13. Shih, T.K., Tang, N.C., Hwang, J.-N.: Exemplar-based Video Inpainting without Ghost Shadow Artifacts by Maintaining Temporal Continuity. *IEEE Trans. on Circuits and Systems for Video Technology* 19(2) (March 2009)

A Coordinate Transformation System Based on the Human Feature Information

Shih-Ming Chang^{1,3}, Joseph Tsai¹, Timothy K. Shih², and Hui-Huang Hsu¹

¹ Department of Computer Science and Information Engineering,
Tamkang University, Taipei County, 25137, Taiwan

² Department of Computer Science and Information Engineering,
National Central University, Taoyuan County, 32001, Taiwan
rest306@hotmail.com

Abstract. In this paper, we propose a method to find feature in human object that used SURF algorithm, and use this information into 3D coordinate that use coordinate system transformation. In our method, first we use thinning algorithm to obtained skeleton of object, and find the endpoints in skeleton. In the second step, we try to use those endpoints to cluster skeleton, and the part number of cluster is six. Then, to cluster human object that use cluster skeleton result. Third, we use SURF algorithm to find the feature in each part in the cluster object image. In this step, we also use SAD method to ensure are correct of feature points that after SURF algorithm treatment. Finally we use the coordinate system transformation method. In this step, we use image coordinate system into world coordinate system, and show those result in our experiments result.

Keywords: SURF, feature points match, object skeleton.

1 Introduction

3D reconstruction technology is a very import in image processing. Until to current, there are many 3D reconstruction methods proposed. In those method [4], there were most used a lot of multi-view images and projection intersection method to build up a 3D model. Then, give mesh on the 3D model surface and give the color information in the mesh path [1]. But this method often had to spent a lot of time in this process. In those methods, they also used some object tracking and feature points matching technology. Those technology can identification object that we want to reconstruction it. Until to current, there are many ready-made tools apply in the 3D reconstruction. Those tools often can produce a good result in reconstruction and object motion, but they also spend a large cost for those tools.

Feature point matching is a very import in image processing. In most image processing, they only want to obtained objects information in the image. If the photography environment is designed, it can obtained object information easy, but if object change the location or angle in photography, it is not easy to detection the feature of object. Until to current, there are many method proposed to solve this

problem. One of those method is Scale-invariant feature transform (SIFT) algorithm [8, 9, 6]. SIFT algorithm is a famous algorithm to find the object's features. It can to detect and describe local feature in image and to find the some features in difference images. In SIFT algorithm, this method used Difference of Gaussian (DOG) function and image pyramid technology to find extreme value in difference scale-space. Then, this algorithm used a linear least square solution and threshold value to decide height-contrast feature points or to excise low-contrast feature points. And use each feature point's gradient direction and feature points strength to allocate of feature points. Final, use histograms to computed orientation value of samples in difference scale-space and create description of the direction of feature points. Because SIFT algorithm can find a large of feature point of object, therefore it can use these feature points to find object in difference image.

One of those methods, Speeded Up Robust Features (SURF) algorithm is also a famous algorithm [2, 3]. SURF algorithm is based on sums of approximated Haar wavelet responses. In the feature described, SURF algorithm is similarity of SIFT algorithm, in the part of detection feature, SIFT algorithm is good than SURF algorithm, but in actual the numbers of feature by use these two methods do not differ too much. In the process time, SURF algorithm is better than SIFT algorithm.

In order to reduce the compute time and more better of feature matching result, there are also many improvement from those feature matching technology. For example, because the SIFT and SURF algorithm maybe to detection wrong feature in object. So, there were proposed method use the segmentation region or Sum of Absolute Difference (SAD) method to reduced the search time and ensure the feature is correct [7]. Used those accessory method, it can effective to reduce the compute time in all procedure and more better in results.

In recent years, coordinate system transform is an interesting research. Until to current, there are many methods and theories that research relationship in multi-coordinate system [10, 11]. In the virtual 3D space, there have some robust methods and theories that transform 3D coordinate into 2D coordinate and show on the screen. It is easy than 2D coordinate into 3D coordinate, because in 2D into 3D, it is only need to reduce a dimensions, and it is a difficult task in 2D into 3D. Because, It is difficult to obtain 3D information from 2D information. Until to current, there are many methods and theories that research how to use multi-view image to obtain 3D information, and transform the 2D coordinate into 3D coordinate.

In this paper, we will use a robust method of feature point matching and an effective transform of image coordinate system into world coordinate system (2D coordinate into 3D coordinate). The section 2 is several processes about thinning objects and find feature in object. Section 3 is several processes about coordinate system transform. Section 4 is our experiment results and section 5 is our conclusion.

2 Feature Points and Skeleton

In our method, we used a simple concept that only used SURF algorithm [2, 3] to find the feature in two object image of difference angle. The result is show in the Fig. 1. In Fig. 1, the result of feature match in the two difference angle images are good. But

there has some match line are matching in wrong feature point, and this result also show a hidden fault that some part have no enough feature points. For example, there are no enough feature points in human object's body and left hand. In order to solve this problem, we used a simple concept to cluster human object and do the SURF algorithm again.

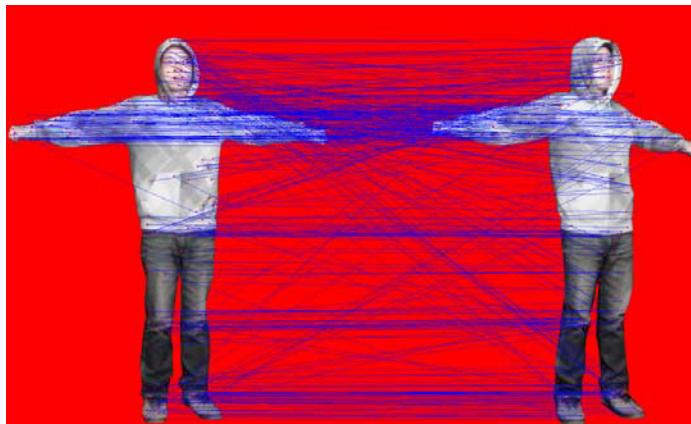


Fig. 1. This figure shows the match of Feature points by SURF

2.1 Find Endpoints in Human Object

We suppose the cause of some part in human object has no enough feature points that the human object is too large lead to the result of feature point maybe to find the similarity but not true feature points. So, we think it can subdivided human object in image.

Until to current, there are many object cluster methods proposed. Most of those methods, they have to spent large time in process. So, we used a thinning algorithm [5] to obtained the skeleton of human object in image, and find the endpoint of this skeleton. In this part, we used three steps to find the endpoints. Step 1 is binary image transform, this method is often used is image process and it can to discriminate object and background. Step 2 is thinning algorithm, this algorithm can find the skeleton from object in image. Step 3 is 5x5 block and eight direction search, and we limit the endpoints that point connected to at least three directions. Those steps are show in follow:

The steps of Find Endpoints:

(1).Load Image.

(2).Do **Binary Image Transform.**

Human object→Pixel transform Black.

Background→Pixel transform White.

(3).Do **Thinning Algorithm** to Obtain Skeleton of Human Object.

(4).Do **5x5 Block and Eight Direction Search.**

if A Point Connected to at Least Three Direction

```

This Point is Endpoint.
Pixel Transform Red.
else
This Point is not Endpoint.
Pixel is Black.

```

(5).End *Find Endpoints*

After these steps, we can obtain the result like Fig. 2. Fig. 2(a) is the skeleton of human object in the image. Fig. 2(b) is show the endpoints, the red points in Fig. 2(b) represent the endpoint in the object skeleton.

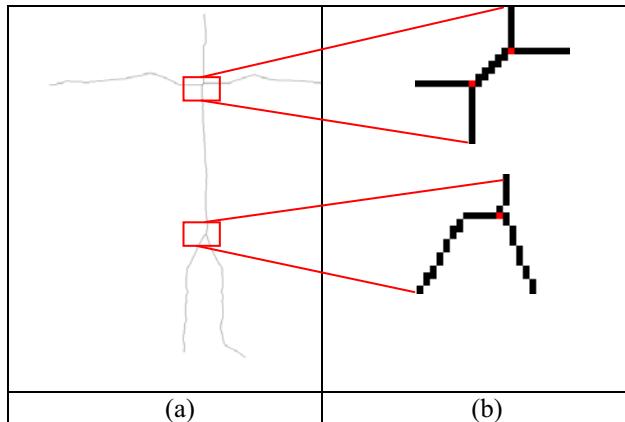


Fig. 2. This figure shows result of Human skeleton and endpoints

2.2 Cluster Object by Endpoints

In Section 2.1, we can obtained the endpoints in the skeleton of human object in image. In one skeleton, it can produced three endpoints, and we can cluster object by those endpoints. We cluster six part of human object skeleton that head, body, left hand, right hand, left foot and right foot, respectively. And give difference color in different part. For an example, we suppose the endpoint is ET1, ET2 and ET3 that order from top to bottom of the red points in Fig. 2(a). In the ET1, we to find the skeleton points by up search. The skeleton points in these rang, we transform color is green, and they are represent the head. In ET3, we find the skeleton points by left search. The skeleton points in these rang, we transform color is light blue, and they are represent the left foot. After the skeleton cluster, we used this skeleton to cluster the object image. In cluster object image, we to find the shortest Euclidean distance ED between object points and difference part of skeleton that difference color. The shortest Euclidean distance is by

$$ED = \sqrt{\text{abs}(\text{SK}_x - \text{OB}_x)^2 + \text{abs}(\text{SK}_y - \text{OB}_y)^2} \quad (1)$$

Which SK_x and SK_y are represent the coordinate (x, y) of the skeleton in image, respectively. OB_x and OB_y are represent the coordinate (x, y) of human object in image. For each pixel of skeleton point have one color, and each pixel of object point transform corresponding color by shortest Euclidean distance ED. Those steps are show in follow:

The steps of Cluster Human Object:

- (1). Load Skeleton Image.
- (2). Load **Endpoints Coordinate** Information.
- (3). **Cluster Skeleton by Endpoints.**
 - (a). **Head:** Up Search from ET1 and Give Green Color.
 - (b). **Body:** Down Search from ET1 to ET3 and Give Blue Color.
 - (c). **Right Hand:** Right Search from ET1 and Give Purple Color.
 - (d). **Left Hand:** Left Search from ET2 and Give Yellow Color.
 - (e). **Right Foot:** Right Search from ET3 and Give Gray Color.
 - (f). **Left Foot:** Left Search from ET3 and Give Light Blue Color.
- (4). Load Human Object Image.
- (5). To Compute Distance for each Object Points by **Formula (1)**.
- (6). To **Give Corresponding Color** by **Shortest ED Value**.
- (7). **End Cluster Human Object.**

After step (3) we can obtained the result of cluster skeleton image like Fig. 3(a), and after step (6) we can obtained the result of cluster human object image like Fig. 3(b).

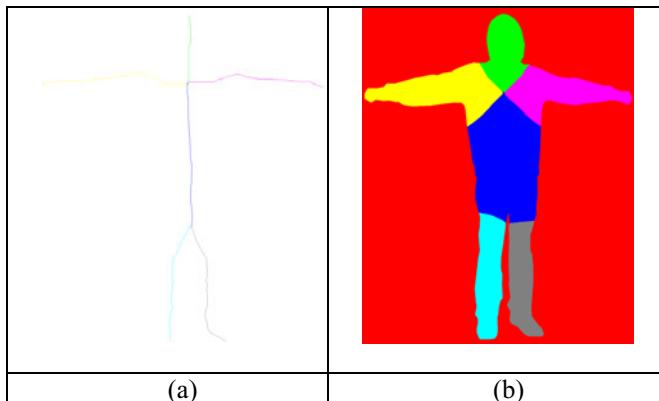


Fig. 3. This figure shows result of cluster human object

Because the human object can be segmentation by a simple concept. Then, we used SURF algorithm [2, 3] again by corresponding part that difference cluster in human object in image. We also use block Sum of Absolute Difference (SAD) base method to ensure the feature points matching are proper [7], and block size in our method is 15x15.



Fig. 4. This figure shows result of cluster object method

The result shows in Fig. 4. In Fig. 4, the angle difference of two input image is 10 degree, and in this figure the accuracy of feature points matching is better than Fig.1 that used SURF algorithm at first.

3 Coordinate System Transform

3.1 Find Camera Parameter Matrix

After we obtained the feature points of human object in image, those feature points can transform in 3D space. In some of traditional method [10, 11], those methods often used projection intersection. But, this method is not appropriate in our research, because the method of projection intersection often used in a complete object and it also to spent large time in this process. In our research, we only used some feature points not used a complete, therefore, we can use coordinate system transform. In this method, feature points can transform into 3D coordinate from the 2D coordinate.

In the coordinate system transform, we have to calibrate camera and find the matrix of camera parameter. We suppose arbitrary feature point in image has a coordinate (u, v) , and the corresponding world coordinate of this feature point is (X_w, Y_w, Z_w) , the relationship of image coordinate system and world coordinate system is:

(2)

Z is an arbitrary multiple. $PM_{11} \sim PM_{34}$ is the matrix of camera parameter, it also can be represent by PMC . By the formula (2), we can expand three simultaneous equations:

(3)

In formula (3), these simultaneous equations can simplify by:

(4)

Because (u, v) and (X_w, Y_w, Z_w) are known, and in formula (4) there have 11 matrix elements is unknown, and the matrix of camera parameter is not 0, therefore, let $PM_{34}=1$ and it need six groups of calibrate points to solve these matrix elements. In this step, we can use the minimal square error method to solve these matrix elements. Therefore, the formula (4) can rewrite by:

(5)

In formula (5), n is represent the number of groups of calibrate points. It also can represent by:

$$P_A * PM_c = Q \quad (6)$$

P_A is the relationship matrix between world coordinate and image coordinate. PM_c is the matrix of camera parameter. Q is the matrix of calibrate points. And it can make use of the minimal square error method to solve PM_c by:

$$PM_c = (P_A * P_A^T)^{-1} * P_A^T * Q \quad (7)$$

Final, we can use formula (2) to formula (7) to solve another matrix of camera parameter, and make use of these camera parameter matrixes to obtain 3D coordinate at follow section. In our research, we used same camera to obtained image that difference angle of human object.

3.2 Compute 3D Coordinate

In section 3.1, we obtained the two matrixes of camera parameter. Therefore, we make use of these camera parameter matrixes and a corresponding point in two input image to compute the corresponding world coordinate. The relationship between camera parameter matrixes and world coordinate is:

$$\mathbf{PM} * \mathbf{W} = \mathbf{N} \quad (8)$$

\mathbf{PM} and \mathbf{N} are the relationship matrixes between camera parameter matrixes and corresponding points. \mathbf{W} is the matrix of world coordinate. Therefore, \mathbf{PM} matrix and \mathbf{N} matrix are known, \mathbf{W} matrix is unknown. It can represent by:

$$(9)$$

(u_1, v_1) and (u_2, v_2) are represent corresponding point in image, respectively. $\mathbf{PM}_{c11}^1 \sim \mathbf{PM}_{c34}^1$ are represent the elements in camera parameter matrix of camera 1, and $\mathbf{PM}_{c11}^2 \sim \mathbf{PM}_{c34}^2$ are represent the elements in camera parameter matrix of camera 2. (X_w, Y_w, Z_w) is represent world coordinate of these corresponding points in 3D space. In the formula (9), it also can make use of minimal square error method to obtain the world coordinate (X_w, Y_w, Z_w) by:

$$\mathbf{W} = (\mathbf{PM}^T * \mathbf{PM})^{-1} * \mathbf{PM}^T * \mathbf{N} \quad (10)$$

Final, we can make use of formula (8) to formula (10) to solve the world coordinate by used camera parameter matrix and a group of corresponding point.

4 Experiments Result

In experiments, the dimension of input image that we used is 1296x2034. The angle of human object images are difference 10 degree that we obtain from same camera. The number of feature point is about 200 that used our proposed method, and those feature points almost all of the correct corresponding in images.

We show some result of feature point matching that make use of our proposed method. The accuracy of those feature points matching is better than only make use of SURF algorithm.

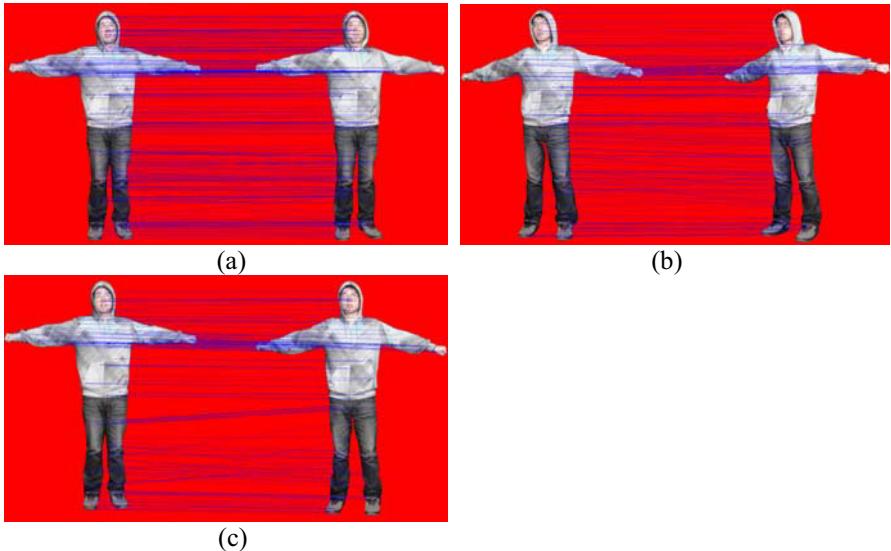


Fig. 5. This figure shows result of feature points matching

In Fig. 5(a), (b), and (c), the angle of these images are difference 10 degree. Because the angle have some difference in these image, the human object have some displacement in the image. But, the feature point matching result impact will not be a large.

We also show some of the 2D coordinate into 3D coordinate in the follow:

Result of image coordinate into world coordinate						
u ₁	v ₁	u ₂	v ₂	X _w	Y _w	Z _w
634	549	670	563	-89.81	-83.88	0.42
621	450	630	446	87.33	-93.85	-1.96
729	610	721	611	-31.96	-100.47	-1.53
694	1211	696	1221	-41.46	21.01	-0.22
629	1410	546	1427	-25.14	-36.97	0.12
703	1502	700	1528	-24.93	-41.35	0.36
561	1416	615	1391	-20.42	-46.08	0.17
739	1551	733	1548	10.65	-164.82	-1.94

Fig. 6. This figure shows result of Coordinate System Transform

In Fig. 6, we can effective transform 2D coordinate into 3D coordinate, and those 3D coordinate can projection in 3D space. Those points in 3D space also can help our

analysis some information like join points in 3D object and let it can be move by those join points.

5 Conclusion

In this paper, we proposed an improved method of SURF algorithm that used with the concept of cluster object, and proposed two algorithm steps to achieve the feature points matching that use thinning algorithm and cluster thinning skeleton to cluster human object. In our method, we don't used traditional method of projection intersection, we used feature points that on the human object in the image, and make use of coordinate system transform to produce the world coordinate of corresponding feature point in human object. The advantage of our method is that the time of process spent less than used projection intersection method. Therefore, in the efficiency of process is better than traditional method.

We also used an effective of coordinate system transform, and described in detail how to use those methods and organized into some simple formula for this method. In the experiments result, we prove that our concept although simple, but it is correct in our method. Then, we also show our result that about feature points matching and 2D coordinate into 3D coordinate.

Final, in the future, we will use the feature points in our method to define some explicit point of human object, it maybe is a key point in the skeleton, then used those key point and skeleton to achieve a 3D skeleton. Therefore, it can achieve an virtual actors in the 3D space.

References

1. Venkatesh, A., Cokkinides, G., Sakis Meliopoulos, A.P.: 3D-Visualization of Power System Data Using Triangulation and Subdivision Techniques. In: Proceedings of the 42nd Hawaii International Conference on System Sciences, pp. 1–8 (2009)
2. Bay, H., Tuytelarrs, T., Gool, L.J.V.: SURF: Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
3. Bay, H., Ess, A., Tuytelarrs, T., Gool, L.J.V.: Speeded-Up Robust Features (SURF). In: Computer Vision and Image Understanding, vol. 110, pp. 346–359 (2008)
4. Matsuyama, T., Wu, X., Takai, T., Nobuhara, S.: Real-time 3D shape reconstruction, dynamic 3D mesh deformation, and high fidelity visualization for 3D video. In: Computer Vision and Image Understanding, vol. 96(3), pp. 393–434 (2004)
5. Zhang, T.Y., Suen, C.Y.: A fast parallel algorithm for thinning digital patterns. Communications of the ACM 27(3), 236–239 (1984)
6. Fazli, S., Pour, H.M., Bouzari, H.: Particle Filter based Object Tracking with Sift and Color Feature. In: International Conference on Machine Vision, pp. 89–93 (2009)
7. Vassiliadis, S., Hakkenes, E.A., Wong, J.S.S.M., Pechanek, G.G.: The Sum-Absolute-Difference Motion Estimation Accelerator. In: 24th. EUROMICRO Conference, vol. 2, pp. 559–566 (1998)
8. Zhao, W.-L., Ngo, C.-W.: Scale-Rotation Invariant Pattern Entropy for Keypoint-Based Near-Duplicate Detection. Image Processing of IEEE Transactions 18(2), 412–423 (2009)

9. Lu, Y., Wang, L., Hartley, R., Li, H., Shen, C.: Multi-view Human Motion Capture with an Improved Deformation Skin Model. In: Computing: Techniques and Applications (DICTA) Digital Image, pp. 420–427 (2008)
10. Zhang, Z.: Flexible Camera Calibration By Viewing a Plane From Unknown Orientations. In: International Conference on Computer Vision, pp. 666–673 (1999)
11. Zhang, Z.: A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(11), 1330–1334 (2000)

An Effective Illumination Compensation Method for Face Recognition

Yea-Shuan Huang and Chu-Yung Li

CSIE Department, Chung-Hua University, Hsinchu, Taiwan
707, Sec.2, WuFu Rd., Hsinchu, Taiwan 300, R.O.C.
yeashuan@chu.edu.tw

Abstract. Face recognition is very useful in many applications, such as safety and surveillance, intelligent robot, and computer login. The reliability and accuracy of such systems will be influenced by the variation of background illumination. Therefore, how to accomplish an effective illumination compensation method for human face image is a key technology for face recognition. Our study uses several computer vision techniques to develop an illumination compensation algorithm to processing the single channel (such as grey level or illumination intensity) face image. The proposed method mainly consists of four processing modules: (1) Homomorphic Filtering, (2) Ratio Image Generation, and (3) Anisotropic Smoothing. Experiments have shown that by applying the proposed method the human face images can be further recognized by conventional classifiers with high recognition accuracy.

Keywords: Face Recognize, Illumination Compensation, Anisotropic Smoothing, Homomorphic Filtering.

1 Introduction

In recent years, digital video signal processing is very popular because digital audio and video technology have made a lot of progress, the price of large data storage is lower and the cost of the optical photographic equipments also decreases. Most importantly, artificial intelligence and computer vision technology are getting mature. So intelligent video processing systems gain much attention to the public, especially it has become a very important role in the safety monitoring field. In this field, the accuracy of face recognition is an essential goal to pursue, so we address this issue here, and hope to be able to develop a high accuracy of face recognition.

For face recognition, there are several problems which will affect the recognition accuracy. Among them, ambient lighting variation is a very crucial problem because it will affect the system performance considerably. Currently, most face recognition methods assume that human face images are taken under uniform illumination, but in fact the background illumination is usually non-uniform and even unstable. Therefore, the face images of the same person often have very different appearances which make face recognition very difficult. Furthermore, slanted illumination probably produces different shadows on face images which may reduce the recognition rate greatly. So

this research focuses on this topic and proposes an illumination compensation method to improve the recognition accuracy under different background illumination.

There are many approaches have been proposed already, such as Retinex [1], Illumination Cone [2], Quotient Image [3], Self-Quotient Image [4], Intrinsic Illumination Subspace [5] , Columnwise linear Transformation [6], Logarithmic Total Variation model [7] , Discrete Cosine Transform [8] algorithm and Gradient faces [9] method. Retinex is an algorithm to simulate human vision which main concept is the perception of the human eye will be affected by the object reflectance spectra and the surrounding lighting source. Therefore, in order to get the ideal image it computes each pixel's albedo by subtracting the intensity of this pixel and those of its surrounding eight pixels, which results in the original Retinex algorithm, also called Single Scale Retinex, SSR. In recent years, several algorithms based on this concept but using more neighboring pixels also were proposed and they proclaimed to produce better performance than Retinex, just like Multi-Scale Retinex, MSR [10] and Multi-Scale Retinex with Color Restoration, MSRCR [10]; Illumination Cone constructs a specific three-dimensional facial model for each person, then various illuminated two-dimensional images of one person can be constructed from his own three-dimensional facial model. All of Quotient Image, Self-Quotient Image and Intrinsic Illumination Subspace adopt an image preprocessing. Quotient Image (QI) has to input at least three images under different background illumination in order to remove the information of lighting source. Self-Quotient Image (SQI) is derived from Quotient Image and it needs only one input image to perform lighting compensation. Therefore, it is easily applied to all kinds of recognition systems. Being similar to QI and SQI, Intrinsic Illumination Subspace first uses a Gaussian Smoothing Kernel to obtain the smoothed image, and then it reconstructs an image with the basis of the intrinsic illumination subspace. Columnwise linear Transformation assumes that by accumulating each column of each human face image the intensity distributions of different persons are very similar. So, the average intensity distribution A nontrivial is computed from all the training face images first, and the intensity distribution B of the current processed face image is also computed, then by transforming B to A, a compensated face image can be derived. The Logarithmic Total Variation (LTV) model is derived from the TV-L¹ model [11] and the TV-L¹ model is particularly suited for separating “large-scale” (like skin area) and “small-scale” (like eyes, mouth and nose) facial components. So the LTV model is also retain the same property. The Discrete Cosine Transform (DCT) algorithm transforms the input image from spatial domain to frequency domain first. Finally, Gradient faces method use Gaussian kernel function to transform the input image to gradient domain and get the Gradient faces to recognition.

However, these methods still have their shortcomings and deficiencies. For example, both Illumination Cone and Quotient Image require several face images of different lighting directions in order to train their database; all of Retinex, Self-Quotient Image, Intrinsic Illumination Subspace , Columnwise linear Transformation, LTV model, DCT algorithm and Gradient faces method cannot tolerate the face angle deviation and certain coverings (such as sunglasses) on faces. For the above reasons, our approach references the previous approaches to propose a novel illumination compensation method. The proposed method is based on “combination” and “complementarity” two key ideas to combine three distinct illumination compensation methods. It can efficiently eliminate the effect of background lighting change, so a

subsequent recognition system can accurately identify human face images under different background illumination.

This paper is arranged into 4 sections. Section 2 describes the concept and the processing steps of the proposed compensation algorithm; Section 3 describes the testing database and experimental results; finally, conclusion is drawn in Section 4.

2 The Proposed Illumination Compensation Method

In order to eliminate the effect of background lighting, we assume that (x, y) is the coordinate of an image pixel P , $f(x, y)$ is the gray value of P . So based on a Lambertian model [12], $f(x, y)$ can be expressed by the multiplication of two functions [2, 3, 12], which is

$$f(x, y) = i(x, y)r(x, y). \quad (1)$$

In this function, $i(x, y)$ is the illuminance of P and $r(x, y)$ is the reflectance of P . In general, the illumination values of neighboring pixels are similar to each other, so $i(x, y)$ can be regarded as one kind of low-frequency signal in an image. However, the reflectance will show the contrast arrangement of different composite materials (such as skin, eyebrows, eyes and lips, etc.) of this image. Therefore, $r(x, y)$ can be regarded as a high-frequency signal which closely corresponds to texture information of face.

Based on this understanding, our research uses the digital filtering approach to reduce the low frequency signal, and emphasize the high frequency signals of a face image at the same time. We expect to decrease the influence of background lighting on facial analysis and recognition. So the facial texture features can be strengthened to achieve the better face recognition accuracy.

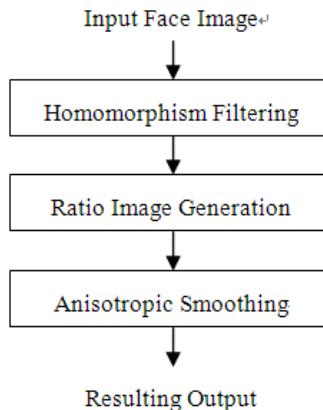


Fig. 1. The processing diagram of the proposed method

The proposed illumination compensation method consists of (1) Homomorphic Filtering, (2) Ratio Image Generation, and (3) Anisotropic Smoothing, which are shown in Fig.1.

2.1 Homomorphic Filtering

In reality, face images are influenced to many conditions and factors (such as lighting and face angle), so an original image may contain lot of noises. Therefore, we use a homomorphic filtering to adjust the image intensity by strengthening the high-frequency signal and decreasing the low-frequency signal.

First, we adopt the logarithm operation to separate the illumination and reflection coefficient from image, that is,

$$\begin{aligned} Z(x, y) &= \ln f(x, y) \\ &= \ln i(x, y) + \ln r(x, y) \end{aligned} \quad (2)$$

Next, we adopt the Fourier Transform to compute the left and right sides of the above equation,

$$\begin{aligned} F\{Z(x, y)\} &= F\{\ln i(x, y)\} + F\{\ln r(x, y)\} \\ Z(u, v) &= F_i(u, v) + F_r(u, v) \end{aligned}$$

where $Z(u, v)$, $F_i(u, v)$ and $F_r(u, v)$ are the Fourier Transform results of $Z(x, y)$, $\ln i(x, y)$ and $\ln r(x, y)$ respectively. Then, we use a low-frequency filtering function $H(u, v)$ to multiply the above formula and get

$$\begin{aligned} S(u, v) &= H(u, v)Z(u, v) \\ &= H(u, v)F_i(u, v) + H(u, v)F_r(u, v) \end{aligned} \quad (3)$$

Furthermore, we use an inverse Fourier Transform to get

$$\begin{aligned} SS(x, y) &= F^{-1}\{S(u, v)\} \\ &= F^{-1}\{H(u, v)F_i(u, v)\} + F^{-1}\{H(u, v)F_r(u, v)\} \\ &= i'(x, y) + r'(x, y) \end{aligned} \quad (4)$$

where

$$\begin{aligned} i'(x, y) &= F^{-1}\{H(u, v)F_i(u, v)\} \\ r'(x, y) &= F^{-1}\{H(u, v)F_r(u, v)\} \end{aligned}$$

Finally, we apply the exponential operation to the above formula and obtain

$$\begin{aligned} g(x, y) &= e^{SS(x, y)} \\ &= e^{\{i'(x, y) + r'(x, y)\}} \\ &= e^{i'(x, y)}e^{r'(x, y)} \\ &= i_0(x, y)r_0(x, y) \end{aligned} \quad (5)$$

After performing all of the above steps, $g(x, y)$ is the final filtered image. Because of $H(u, v)$ is a low-frequency filtering function, it will significantly reduce the intensity of low-frequency signal. So $g(x, y)$ can not only effectively preserve the high-frequency texture information, but also reduce the impact of illumination variation.

In general, $H(u, v)$ can be designed as

$$H(u, v) = (r_H - r_L)[1 - e^{-C[D^2(u, v)/D_0^2]}] + r_L \quad (6)$$

where $r_H > 1$, $r_L < 1$, and D_0 is a cut-off frequency. The constant c is a parameter to control the increasing degree of the exponential function. Figure 2 shows an illustrating graph of $H(u, v)$.

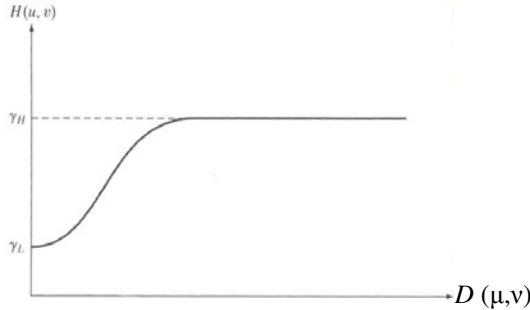


Fig. 2. A low-frequency filtering function $H(u, v)$

The low-frequency signal not only includes the illumination information but also includes the texture information of human face image. So the r_L should be set to a small value but not zero, if we want not destroy the texture information of face image. Because of above reason, in order to remove the illumination information, we proposed the second steps: Ratio Image Generation.

2.2 Ratio Image Generation

We have used the homomorphic filter to reduce the influence of illumination, but we cannot eliminate all low-frequency signals because the low-frequency signal may also contain some facial features which are useful to recognition. So instead of setting $r_L = 0$ to completely eliminate the low-frequency signal, r_L is set to be 0.5. Consequently, the filtered image still contains part of illumination information. For further reducing the illumination information, a second operation called “Ratio Image Generation” is proposed to eliminate the low-frequency signal. From the experiment, it clearly shows that using both of Homomorphic Filtering and Ratio Image Generation outperform than using Homomorphic Filtering only.

Since $g(x, y)$ denotes the value of a filtered image pixel, based on a Lambertian model [8], it can also be formulated as

$$g(x, y) = r(x, y)i(x, y) \quad (1)$$

where $r(x, y)$ is the albedo, and $i(x, y)s(x)$ is the illumination value of pixel (x, y) . As described before, $r(x, y)$ denotes the texture information of the image and $i(x, y)$ denotes the low-frequency information. Let $W(x, y)$ be a smoothed image information by convoluting $g(x, y)$ with a Gaussian function G . That is

$$W(x, y) = g(x, y) * G. \quad (2)$$

Basically, the lighting factor can be implicitly attributed to W . Because both $i(x, y)$ and $W(x, y)$ correspond to the low frequency signal of an image at pixel (x, y) , we can

use $i(x, y) \approx c \times W(x, y)$ to present the approximate relationship between both low-frequency data and c is a constant value. If $g(x, y)$ is divided by $W(x, y)$, a new image $N(x, y)$ can be constructed which inherently reveals the high frequency attribute $r(x, y)$. That is

$$N(x, y) = \frac{g(x, y)}{W(x, y)} = \frac{r(x, y) i(x, y)}{W(x, y)} \approx cr(x, y) \quad (3)$$

where N can effectively reflect the intrinsic information of an image, which is called the ratio image.

2.3 Anisotropic Smoothing

While a ratio image N can effectively reflect the high-frequency signal of image, but it is very sensitive to noise. Therefore, we use an anisotropic smoothing operation to reduce the interference of noise. However, the general smoothing algorithms will not only reduce noise, but also undermine the image texture characteristics because they belong to high frequency signal. In order to reduce the noise effect and avoid the degeneration of normal texture information, we purposely design an anisotropic smoothing algorithm to produce the smoothed image. Here, some variables about the anisotropic smoothing operation are defined as below:

$N_{x,y}$ is the image value of pixel (x, y) in a ratio image

$$\begin{aligned} \Delta_E &= N_{x+1,y} - N_{x,y} \\ \Delta_W &= N_{x-1,y} - N_{x,y} \\ \Delta_S &= N_{x,y+1} - N_{x,y} \\ \Delta_N &= N_{x,y-1} - N_{x,y} \end{aligned} \quad (4)$$

Δ_E , Δ_W , Δ_S and Δ_N represent respectively the 4-directional image differences between pixel (x, y) and its adjacent image pixels. During the smoothing operation, a large degree of smoothing will be executed on the uniform parts of image, but a much small degree of smoothing will be executed on the boundary of image. Consequently, the smoothed image will preserve its boundary information effectively. To serve this purpose, a weighting function based on image difference is designed as

$$w_k = \exp^{-\frac{\Delta_k \cdot \Delta_k}{\delta}} \quad \text{for } k \in \{E, W, S, N\} \quad (5)$$

where δ is the bandwidth parameter to control the change rate of the exponential function. Then, the smoothed image are computed by

$$g_{x,y}^t = g_{x,y}^{t-1} + \lambda(w_E g_{x+1,y} + w_W g_{x-1,y} + w_S g_{x,y+1} + w_N g_{x,y-1}) \quad (6)$$

where $g_{x,y}^t$ is the image value of pixel (x, y) after t times smoothing operations. Finally, in order to obtain more consistently filtered face images, a histogram equalization operation is applied to the anisotropic smoothed image.

3 Experimental Results

In order to estimate the performance of the proposed method, the present study uses two famous face databases (Banca [13] and Yale database B [14]) to evaluate the recognition rate. The Banca database contains human frontal face images grabbed from several sections to reflect different variation factors. Among all sections, the section 1, 2, 3 and 4 of the “controlled” classification are used in our experiment. In each section, there are 10 images for each person, and in total there are 52 persons (26 males and 26 females), therefore it consists of 2,080 images in total. For performance comparison, we adopted three pattern matching methods (RAW, CMSM [15] and GDA [16]) to evaluate the recognition accuracy. RAW refers to the nearest-neighbor classification based on the image value in the Euclidean distance metric. CMSM (Constrained Mutual Subspace Method) constructs a class subspace for each person and makes the relation between class subspaces by projecting them onto a generalized difference subspace so that the canonical angles between subspaces are enlarged to approach to the orthogonal relation. GDA (Generalized Discriminant Analysis) adopts kernel function operator to make it easy to extend and generalize the classical Linear Discriminant Analysis to a non-linear one. Because CMSM needs to construct a mutual subspace, the images of 12 persons are selected to serve this end. Therefore, the face images of the rest 40 persons are used to test the recognition performance in this experiment. By randomly separating the 40 persons, different enrollment and unenrollment sets are constructed. An enrollment set contains the face images of the persons which have enrolled themselves to the recognition system and an unenrollment set contains the face images of the persons which have not enrolled to the system. During each random separation, there are 35 persons are selected in the enrollment set and 5 persons are in the unenrollment set. With this design, hundreds of experiments can be easily performed. Among the four sections, only the first section is used for serving the training purpose, and the other three sections are for testing. As for the Yale database B, it contains 5760 single light source images of 10 subjects each was taken pictures under 576 viewing conditions (9 poses x 64 illumination conditions). For every subject in a particular pose, an image with ambient (background) illumination was also captured. Hence, the total number of images is in fact $5760+90=5850$. But we only test 1 pose (pose 0) of them; it means we only use 640 images to test the recognition rate. Then these 64 images are further separated into 6 sections (about 10 images per section), and only the first section is used for serving the training purpose, and the other five sections are for testing. Among 10 peoples, 5 of them are selected for enrollment, and the other 5 are for unenrollment.

The specific settings of parameters in our experiments are $r_H = 1.6$, $r_L = 0.5$, $D_0 = 15$, and $c = 10$. For CMSM, the base number is set 1000, and for GDA, the kernel sigma is set 4400 and its feature dimension is 200. Figure 3 shows some images examples of which the first row is the original images, the second row is the images after applying the homomorphic filter, the third row is the ratio images, and the fourth row is the images operated by the anisotropic smoothing algorithm which indeed are the output images of our illumination compensation method.

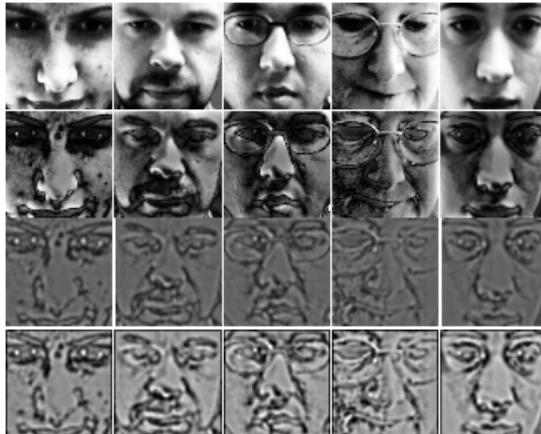


Fig. 3. Image examples of different processing steps, from the first row to the fifth row: input images, hormomorphically filtered images, ratio images, and anisotropic smoothed images

Table 1 lists the recognition results of three different pattern matching methods, and FAR, FRR, and RR denote false accept rate, false rejection rate, and recognition rate individually. From this table, it shows all the recognition rates of the three recognition methods are larger than 90%, and the recognition rate of CMSM even is up to 95%. Thus, this experiment demonstrates that the compensated image by using the proposed approach can be further recognized by general recognition methods.

Table 1. The recognition results of three different pattern matching methods on the compensated Banca face database

	CMSM	RAW	GDA
FAR	4.6%	6.7%	6.1%
FRR	4.8%	7.3%	6.5%
RR	95.1%	92.6%	93.4%

In addition, this study also compared the recognition rates with eight other illumination compensation methods: (1) Original, means we used original image to processing image without illumination compensation, (2) HE, means Histogram equalization method, (3) Retinex, (4) DCT means the Discrete Cosine Transform algorithm, (5) RA, means that we used ratio image generation + anisotropic smoothing, (6) HA is means homomorphic filtering + anisotropic smoothing, (7) LTV means Logarithmic Total Variation model, and (8) Gradient faces method. Besides, the recognition result of the original images is listed as a reference. Table 2 shows the experimental results to compare our algorithm with other compensated algorithms. Obviously, our method outperforms the other methods.

Because the Banca database does not contain images with significant illumination variation, we purposely used a few human face images from the Yale Face database

Table 2. Recognition results of different illumination compensation algorithms adopt Banca database

Illumination Compensation Method	CMSM	RAW	GDA
Original	88.2%	57.6%	60.3%
Histogram equalization	88.5%	60.1%	64.3%
Retinex	81.5%	65.0%	75.4%
DCT	88.3%	85.1%	82.0%
RA	89.1%	88.3%	84.1%
HA	91.8%	85.0%	81.7%
LTV	92.3%	90.0%	90.6%
Gradient faces	92.5%	87.4%	90.7%
The proposed method	95.1%	92.6%	93.4%

[16] to demonstrate the effectiveness of our illumination compensation method. Visually, from Figure 4, the original images in the first row show different appearances, but the final output images in the fourth row in fact appear quite similar to each other. Table 3 lists the recognition rates of our proposed method and the other illumination compensation algorithms with the Yale database B.

**Fig. 4.** Image examples from Yale faces database. The first column to the third column is respectively “central-light source image”, “left-light source image”, and “right-light source image”.

Table 3. Recognition results of different illumination compensation algorithms and different databases

Illumination Compensation Method	CMSM	RAW	GDA
Original	90.0%	82.6%	92.2%
Histogram equalization	96.1%	91.6%	97.9%
Retinex	95.8%	87.8%	97.8%
DCT	94.0%	88.0%	100.0%
RA	93.9%	83.7%	95.9%
HA	92.1%	87.6%	97.8%
LTV	86.2%	93.0%	98.2%
Gradient faces	94.1%	93.8%	100.0%
The proposed method	97.8%	95.6%	100.0%

In Table 3, We can find the recognition rate of Yale database B can be up to 100%. It's because the Yale database B contains variation only in illumination and keeps other conditions (ex. background, pose, expression and accessory) the same. However, the recognition rate of Banca database is lower (at most 95.1%) because it basically contains more variation and more peoples than Yale database B. So we can say the recognition of Banca database is more difficult than that of Yale database B. From the above experiments, it obviously shows that our purposed method consistently performs best than the other commonly used illumination compensation methods for the Banca, and Yale B face databases.

4 Conclusion

In this paper, we propose a set of illumination compensation technique use for human face recognition. The proposed technique uses digital filtering to reduce the low-frequency signal and strengthen the high-frequency signal to reserve the facial texture information. And the proposed technique also can reduce the effect of background lighting change to increase the accuracy of face image recognition. Experiments have shown that the proposed method can achieve very promising recognition accuracy for the Banca database and Yale B faces database of each recognition method. It confirms the proposed algorithm is indeed more feasible and applicable. Actually, the proposed method is a general lighting compensation method which is not only limited in recognizing human faces. In the future, we will try to apply this method to other applications (such as OCR and Video surveillance).

References

1. Lu, C.-L., Wang, Y.-K., Fan, K.-C.: Face Recognition under Illumination and Facial Expression Variation. In: CVGIP (2009)

2. Belhumeur, P.N., Kriegman, D.J.: What is the set of images of an object under all possible lighting conditions? In: Proceedings, IEEE International Conference on Computer Vision and Pattern Recognition, pp. 52–58 (1997)
3. Raviv, T.R., Shashua, A.: The quotient image: Class based re-rendering and recognition with varying illuminations. In: Proceedings, IEEE International Conference on Computer Vision and Pattern Recognition, pp. 566–571 (1999)
4. Wang, H.T., Li, S.Z., Wang, Y.S.: Face Recognition under Varying Lighting Conditions Using Self Quotient Image. In: Proceedings, IEEE International Conference on Automatic Face and Gesture Recognition, pp. 819–824 (May 2004)
5. Chen, C.P., Chen, C.S.: Lighting Normalization with Generic Intrinsic Illumination Subspace for Face Recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (October 2005)
6. Lee, M., Park, C.H.: An Efficient Image Normalization Method for Face Recognition Under Varying Illuminations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 711–720 (1997)
7. Chen, W., Er, M.J., Wu, S.: Illumination Compensation and Normalization for Robust Face Recognition Using Discrete Cosine Transform in Logarithm Domain. IEEE Transactions on Systems, Man, and Cybernetics 36(2) (April 2006)
8. Chen, T., Yin, W., Zhou, X., Comaniciu, D., Huang, T.S.: Total variation models for variable lighting face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(9), 1519–1524 (2006)
9. Zhang, T., Tang, Y.Y., Shang, Z.: Face Recognition Under Varying Illumination Using Gradientfaces. IEEE Transaction on image processing 18(11) (November 2009)
10. Li, Y.-J.: Color Image Enhancement Using Hybrid Retinex Algorithm. Master Thesis of Department of Graphic Communications and Technology (2004)
11. Chen, T.F., Esedoglu, S.: Aspects of Total Variation Regularizes L^1 Function Approximation. CAM Report 04-07, Univ. of California, Los Angeles (February 2004)
12. Oren, M., Nayar, S.K.: Generalization of the Lambertian model and implications for machine vision. International Journal of Computer Vision 14(3), 227–251 (1995)
13. The BANCA Database – English part, <http://banca.ee.surrey.ac.uk/>
14. The Yale faces database,
<http://cvc.yale.edu/projects/yalefaces/yalefaces.html>
15. Giraudon, C.: Optimum antenna processing: a modular approach. In: Proc. NATO Advanced Study Inst. on Signal Processing and Underwater Acoustics, pp. 401–410 (1977)
16. Baudat, G., Anouar, F.: Generalized discriminant analysis using a kernel approach. Neural Comput. 12, 2385–2404 (2000)

Shape Stylized Face Caricatures

Nguyen Kim Hai Le¹, Yong Peng Why², and Golam Ashraf¹

¹ School of Computing, National University of Singapore

² Faculty of Arts and Social Sciences (Psychology), National University of Singapore
`{dcslnkh, psywyp, gashraf}@nus.edu.sg`

Abstract. Facial caricatures exaggerate key features to emphasize unique structural and personality traits. It is quite a challenge to retain the identity of the original person despite the exaggerations. We find that primitive shapes are well known for representing certain personality traits, in art and psychology literature. Unfortunately, current automated caricature generation techniques ignore the role of primitive shapes in stylization. These methods are limited to emphasizing key distances from a fixed Golden Ratio, or computing the best mapping in a proprietary example set of (real-image, cartoon portrait) pairs. We propose a novel stylization algorithm that allows expressive vector control with primitive shapes. We propose three shape-inspired ideas for caricature generation from input frontal face portraits: 1) Extrapolation in the Golden Ratio and Primitive Shape Spaces; 2) Use of art and psychology stereotype rules; 3) Constrained adaptation to a supplied cartoon mask. We adopt a recent mesh-less parametric image warp algorithm for the hair, face and facial features (eyes, mouth, eyebrows, nose, and ears) that provides fast results. The user can synthesize a range of caricatures by changing the number of identity constraints, relaxing shape change constraints, and controlling a global exaggeration scaling factor. Different cartoon templates and art rules can make the person's caricature mimic different personalities, and yet retain basic identity. The proposed method is easy to use and implement, and can be extended to create animated facial caricatures for games, film and interactive media applications.

Keywords: Face caricatures, face cartoon, image warping, shape psychology, cartoon exemplar.

1 Introduction

Face caricatures are commonly used in the print media and animation industry. A caricature usually exaggerates a person's face features by emphasizing unique physical or personality traits. However, exaggeration of unremarkable face features might lead to misrepresentation of the person's identity. Some existing methods have employed identity preservation by unequal scaling of features based on feature distances from a certain Golden Ratio. Others have attempted to learn the caricature rules from examples. None of these methods explicitly formulate the role of primitive shapes, which is fairly evident in stylized caricatures. Primitive shapes are popularly used to represent certain personality traits[3, 21]. Square shapes generally represent stability, boredom, meanness, etc. Round shapes portray happiness, closure and balance. Inverted triangles represent instability, shrewdness, evil, etc.

In this paper, we decompose the human face into a number of component parts, using a recently developed shape representation algorithm [14]. This representation allows a smooth transition between the circle, triangle and square shapes, a property that we will exploit for mesh-less image warping [2] in this paper. Since the widely employed Active Shape Model is incapable of handling feature segmentation of arbitrarily exaggerated cartoon templates, we rely on manual annotation strokes for the segmentation of face component shapes in both cartoons and real faces. These strokes are fitted into mixed primitive vector shapes using least-square error minimization after scale and rotation normalization [14]. Our system accepts a frontal human portrait with neutral expression and transforms it into a caricature (based on shape-vector segmentation of upper face, jaw, eyes, nose, ears, mouth and hair) with one of the following three methods: 1) Extrapolation in the Golden Ratio and Primitive Shape Spaces; 2) Use of art and psychology stereotype rules; 3) Constrained adaptation to a supplied cartoon mask. We adopt a mesh-less image warping algorithm [2] that flexibly scales, rotates, positions and shapes face components, to create the final result.

Our main contribution in this paper is the introduction of mixed and pure primitive shapes, in stylizing the personality and structure of face portraits. We extend salient ideas of selective exaggeration from Gooch et al. [11] in the “Golden Ratio feature-distance scaling” method, and further build on this to propose a novel decomposition of key exaggeration and suppression features, for constrained adaptation to cartoon templates. The rest of the paper is organized as follows. Section 2 has the related work to our paper. In Section 3, we briefly describe our prior work that will play a part in our system. Section 4 and 5 will describe our approach in producing caricatures. In Section 6, we will explain the shape-field image warping method and its multi-pass application for the face.

2 Related Work

Gooch et al. [11] have postulated that humans recognize faces based on the amount that facial features deviate from an average face. They have presented an interactive caricature generator by using a method for creating two-tone illustrations and warping the face using a 3x3 Freeform Deformation grid. They were perhaps the first to stress on the identity issues of caricature generation. In this paper, we propose that identity is not just related to key differences from the mean, but also set up by the regular features that are close to the mean. Chiang et al. [7] have also exaggerated the prominent features of the input face after comparison with the average face. Others [1, 5] have proposed interactive caricature generator tools. Most of these approaches do not attempt to learn from expert artists.

On the other hand, there have been some proposed systems to teach a computer to automatically generate a stylized facial sketch by observing images drawn by artists. [18, 19] have proposed a method that learns to exaggerate features based on physical landmarks. Liang et al. [18] have predicted the best {style, real-face exemplar} pair given an input face, using Partial Least Squares learning. Liu et al. [19] have used PCA reduction to map the input face to a closest neighbor, and thus its corresponding pre-drawn cartoon. Their results do not always produce appropriate identity and shape

exaggeration, even though artist knowledge seems to have been inferred by machine learning. Chen et al. [6] have learned art style from training pairs of images and associated sketches drawn in a particular style. The model generates a stylistic sketch from input images, but does not learn how to exaggerate it. Freeman et al. [9] has presented an example-based system for translating a sketch into different styles. However, they focused on style transfer rather than generating a stylistic sketch from an image.

Anthropometric studies [20, 23] have extensively studied facial proportions in terms of perception of beauty across different countries and races. We draw from the Golden Ratio rules proposed here and use it as our basis of discriminating between exaggeration and suppression features. In this paper, we present three different strategies: 1) to re-visit mean feature exaggeration, similar to [7, 11]; 2) to distort shapes with procedural rules inspired by art, a novel area; 3) to adapt to an existing cartoon as a constrained optimization problem, an extension to Gooch et al.’s idea of identity [11].

3 Our Prior Work

We refer the reader to our prior work in shape representation [14] and image warping [2] as we use it in this paper to characterize and deform face images.

As shown in Fig.1, we store each of the three normalized primitive shapes as a set of eight quadratic Bezier curves. The solid points represent segment boundaries and the ragged blotches represent mid-segment control points. The normalized shapes can be affine transformed to any location, scale and rotation. Each shape vector (henceforth informally termed as *cage*) is compactly represented as $\{w_O, w_\Delta, w_\square, ctr_x, ctr_y, rotation, length, breadth\}$. Results of some blend operations are shown in Fig.1b. The cross hair under the shapes indicates the shape weights.

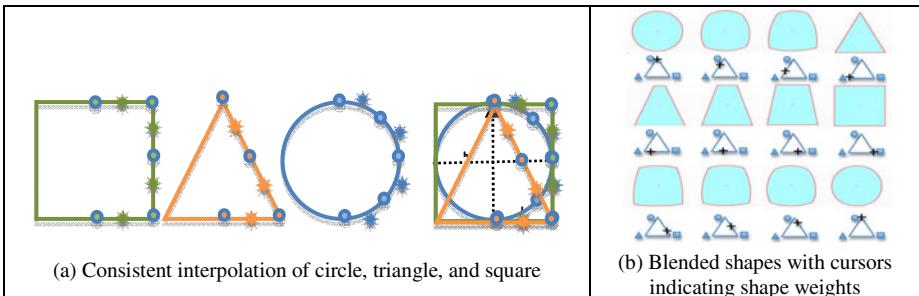


Fig. 1. Piecewise quadratic Bezier construction and blending of shape vector [14]

The above representation performs well for roughly symmetric and convex shapes. This assumption is fairly reasonable for human faces and facial features. Once the facial features are tracked or marked up, they can be vector fitted following a process outlined in [2]. As shown in Fig. 2, we rely on s, t parameterization of the vector cage to perform image warping [2]. Radial parameter t is a real number that also indexes into a lookup table of cage outline points. Parameter s for any given Cartesian point p is its distance from the centroid, scaled by the corresponding center-to-boundary

distance r . To avoid repeated curve intersection calculations, we cache r (360 rows @ 1° increments) for every cage in the scene graph.

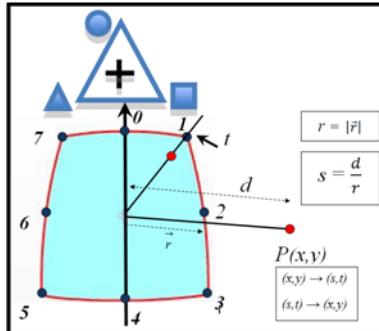


Fig. 2. Polar Coordinate Parameterization of a Cage [2]

Fig. 3 shows the stroke annotation tool that can be used to trace out face components. Each component is later vector fitted using least square error minimization described earlier. The annotation process takes only a few minutes per face. We currently mark the following components: *upper head, jaw, eyebrows, eyes, nose, ears, mouth and hair*. Since hair texture also contributes to the identity of the character, and may affect the result of warping, we construct a hair mask (blue outline in Fig. 3). We will explain more about multi-pass image warping in Sec.6.

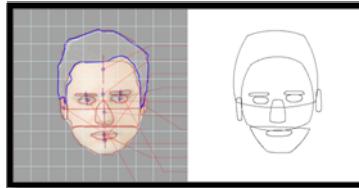


Fig. 3. Face Annotation and Fitted Vector Cages

4 Golden Ratio Feature Space

The variables shown in Fig. 4 were originally derived after performing Principal Component Analysis on extensive anthropometric survey data on different races [20,23]. The rules in Fig. 4 embody a Golden Ratio for beauty, according to benchmarks for the Caucasian race.

We evaluate the LHS of the equations in Fig. 4 from our vector annotation and construct a Golden Ratio Feature Vector $\overline{\mathbf{K}}^{\#} = [k1, k2, k3, k4, k5, k6, k7]$, where each k_i in $\overline{\mathbf{K}}^{\#}$ indicates how different that feature is to the Golden Ratio Mean face. Thus the face vector space S can be thought of as a 7D coordinate system, where real faces lie close to the origin, and exaggerated faces lie further away along one, few or all the feature component vectors.

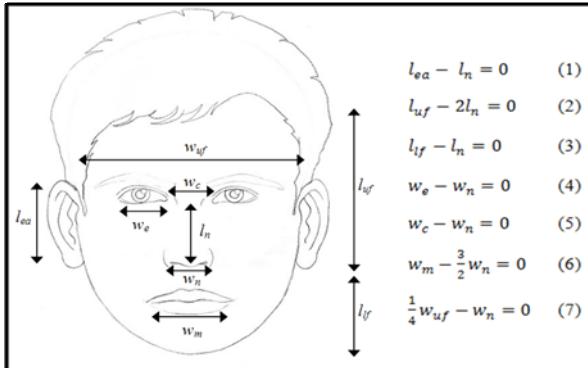


Fig. 4. Golden Ratio Rules [20, 23]

5 Caricature Generation

5.1 Golden Ratio Based

The main idea is to exaggerate feature distances using Golden Ratio, as well as shape. Existing extrapolation methods so far [11] just handle the distances. As shown in Eqn.1, this method simply scales positive or negative offsets in the Golden Ratio Feature Vector of an annotated face, with a suppression offset that clamps features that are very close the mean to preserve identity.

$$k'_i = (1 + \alpha)k_i, \quad \text{if } (k_i > \phi). \quad (1)$$

The shapes are changed according to the following pseudo-code:

```

Exaggeration ratio:  $\alpha$  (0< $\alpha$ <1)
oldShape = [wo, we, wn]
minShape = min(wo, we, wn);
maxShape = max(wo, we, wn);
if (wi is maxShape) // w0=wo, w1=we, w2=wn
    wi += minShape *  $\alpha$ ;
else if (wi is not maxShape and minShape)
    wi += minShape * (1-  $\alpha$ );
else wi = 0;
newShape = [wo, we, wn]

```

Fig.5a shows results of different scales of shape and distance exaggeration. Features closer to the mean are minimally affected by the exaggeration ratio α . Since the input face character has a larger than usual jaw, it is affected the most by exaggeration. Furthermore, the hair and jaw cages undergo a significant shape change towards (\square) and (Δ) respectively, because the base shape is already quite close to these shape-weight edges (see weight space abstraction in Fig.2b). Also, the size of the eyes, ears, and lips were automatically picked up for enlargement by our method.

Fig.5b proves that our method produces better shape stylization than Liu et al. [19], hence retaining better identity of the original person. Their method does not differentiate between expressive and close-to-mean features, resulting in all the components undergoing arbitrarily styled warps. We conducted an online survey for objective comparison. We first retrieved a group of five faces that are quite close to the original face image in Fig 5, using an online face recognition and retrieval system implemented in [24]. We then set up two separate online surveys that required users to match our and Liu et al.'s cartoon to the faces retrieved from [24] (along with the original image in Fig 5). The survey was done with grayscale images to remove skin color related cues. Results from 36 anonymous surfers have been tabulated in Fig. 6.

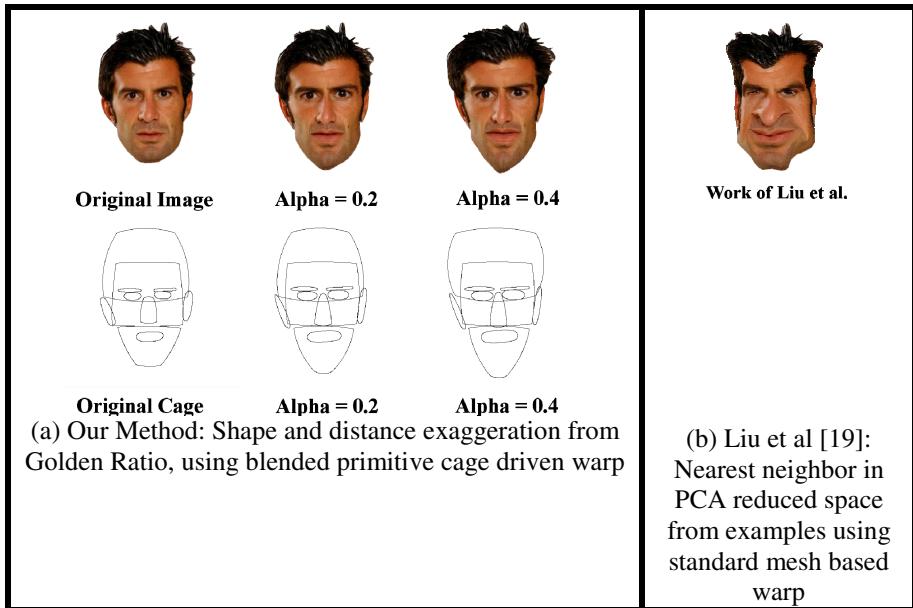


Fig. 5. Comparison of our method with Liu et al [19]. Shape and identity is better preserved and exaggerated in (a) compared to (b).

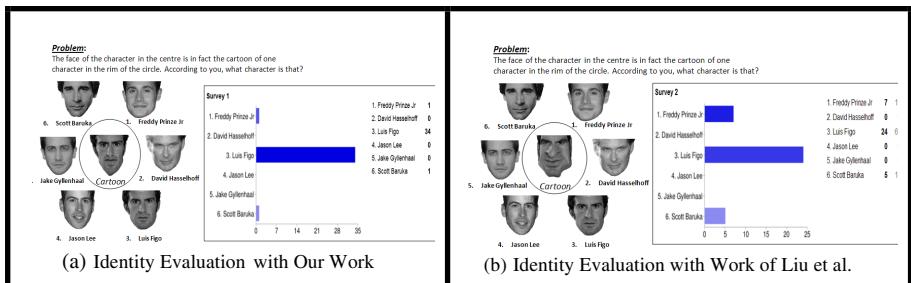


Fig. 6. Comparison of identity survey results

About 94% of the respondents correctly identify the original character using our cartoons, while the number for Liu et al. is 67%. We conducted another survey to let users rate the better caricature (our cartoon in Fig. 5a with $\alpha=0.4$, vs. Liu et al.'s cartoon in Fig. 5b). Our cartoon got favorable votes from 69% respondents. This clearly proves that our method produces better cartoonification without compromising the identity of the character.

5.2 Art and Psychology Stereotype Based

We encountered a few art sources [3, 12] which detail specific shapes for eyes, eyebrows and face shape. We attempted biasing certain primitive shape weights to procedurally stylize a given face image to different stereotypes. The pseudo code for the Innocent, Mean and Devious stereotypes is listed below, and image-warping results are shown in Fig. 7. The exaggeration ratio α was set to 0.2. We find the results quite interesting, and find useful results for a range of exaggeration values and rule variations.

Mean face:	Devious face:	Innocent face:
<ul style="list-style-type: none"> - Square eye: $w_o = 100$; - Square eyebrow: $w_o = 100$; - Thicker eyebrow: $height += height * \alpha$; - Thicker eye: $height += height * \alpha$; - Square and thinner upper nose: $w_o = 100$ $width -= width * \alpha$ 	<ul style="list-style-type: none"> - More triangle and non square eyes: $w_A = 70$; $w_o = 30$ - More triangle and non square lower nose: $w_A = 70$; $w_o = 30$ - More triangle and non square lower face: $w_A = 70$; $w_o = 30$ - Thinner eyebrow: $height -= height * \alpha$ - Less broad face: $width -= width * \alpha$ - Less broad nose: $width -= width * \alpha$ 	<ul style="list-style-type: none"> - Round eye: $w_o = 100$ - Bigger eye: $height += height * \alpha$; $width += width * \alpha$; - Thinner eyebrow: $height -= height * \alpha$; - Round chin: $w_o = 100$; - Round nose: $w_o = 100$;

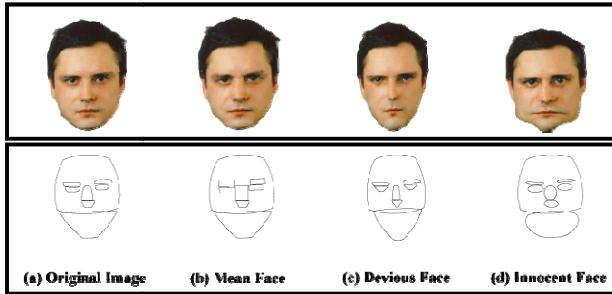


Fig. 7. Procedural primitive shapes for establishing character stereotypes

5.3 Cartoon Template Based

We model the cartoon template adaptation for caricature generation as a two-way linear constraint optimization problem, where the goal is to find the best compromise between exaggeration and identity retention. The pseudo code below describes the

algorithm, in which we first partition the Golden Ratio Mask features of a real face into expressive and suppressive feature sets. The main idea is that we want to exaggerate the real-face features that are far away from the mean, as well as suppress (anti-caricature) those cartoon features whose corresponding real-face features are close to the mean. The best positive and negative exaggeration factors are iteratively searched in a min-max range of possible scales, with a goal to minimize errors with the exaggerated cartoon features and suppressed real features. After these scale factors are extracted, the deformation cages are constructed subject to overlap constraint correction, followed by the image warping.

```

- Read Input face data, read cartoon data
- Construct  $K_i[]$  for input face and  $K_j[]$  for cartoon face.
- Detect Exaggeration Set in input face:
  If ( $|K_{2i}| > |K_{1i}|$ ) and  $K_{2i}$  and  $K_{1i}$  have same polarity:
     $K_i$  in the Exaggeration Set
  Otherwise:  $K_i$  in Suppression Set
- In Exaggeration Set:
  Find  $\alpha^+$  in min-max( $K_{2i} / K_{1i}$ ) to minimize  $\Sigma (|K_{2i} - K_{1i} * \alpha^+|)$ 
- In Suppression Set:
  Find  $\alpha^-$  in min-max( $K_{1i} / K_{2i}$ ) to minimize  $\Sigma (|K_{1i} - K_{2i} * \alpha^-|)$ 
- Construct the Target Vector  $K$ 
  if  $K_{1i}$  in Exaggeration Set:  $K_i = K_{1i} * \alpha^+$ 
  else Suppression Set:  $K_i = K_{2i} * \alpha^-$ 
- Adjust the cages based on:
  Construct cages from  $K_i$  without violating
  overlapping constraints
  Final cage shapes in target model mimic the cartoon model
- Warp Input Image

```

Fig. 8 shows some caricatures of same input faces. Output face caricatures adapt well to the cartoon template, yet still retain the identity of characters. We applied the Caucasian Golden Ratio on the Asian male out of curiosity, but got some interesting results as well. In general, we can see that the Caucasian man's caricatures are much more conservative than the Asian man, since his features are closer to the Male Caucasian Golden Ratio values. This clearly illustrates the expressive power of our approach, where we can get different degrees of exaggeration by varying identity

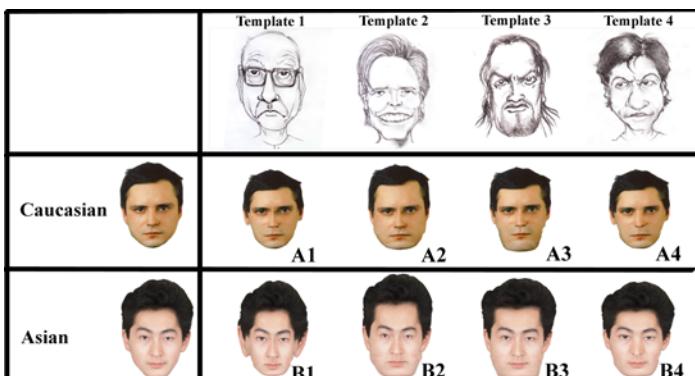


Fig. 8. Caricature generation by template adaptation

constraint thresholds and exaggeration factor α . Fig. 9 illustrates the cartoon identification results of a quick online survey with 36 respondents. Fig. 9a shows a typical survey page, where the respondent needs to match the grayscale cartoon in the center with one of the originals. Fig. 9b illustrates the high percentage (86-97%) of correct identification for all four cartoon versions (B1-B4 in Fig. 8) of the Asian man.

The results of audience correlation to the source templates are tabulated in Fig. 10, with cross marks indicating the correct response for each cartoon. Though the precision is significantly lower (39-58%) than the face identity results, the correct templates still received a clear majority (>10% higher than wrong categories). Some reasons for the observed confusion might be: i) the source templates share certain distinct common shapes (e.g. face outline: Templates 2 and 4, forehead shape: Templates 1, 2 and 3); ii) Certain features may lie in conflicting (i.e. exaggeration and suppression) sets for the source image and cartoon template.



Fig. 9. Online survey for identity evaluation of cartoon template based method

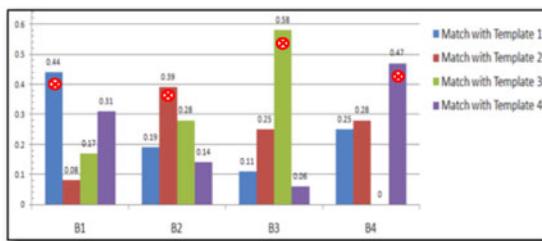


Fig. 10. Results of audience guess for templates used to generate cartoons B1-B4

6 Image Warping

We refer the reader to [2] for details on the parametric mesh-less image-warping algorithm. One of the limitations of mesh based warping methods is that they yield artifacts like folds (slim triangles), stretches and tears (overlapped triangles) when corresponding points in the target move by large distances. Some methods try to limit the amount of deformation to alleviate these problems [13, 25], but caricatures need to

deform significantly. Though this warping method does not guarantee against stretching artifacts, it does improve the fold and tear problem significantly.

We now describe three problems we need to address for face warping. Firstly, the hair cage, upper face cage and the jaw cage themselves envelope other face component cages. Weighted combination of influences is an option, but we need to select the influence factor carefully (which obstructs procedural synthesis). Secondly, a skin pixel not belonging to any sub-components but belonging to the face could end up getting sourced from one of the component locations, due to the weighted influence from multiple cages. Thirdly, the hair cage overlaps a large part of the face, but should not really affect the deformation of the face outline or face components.

In order to address all the above issues, we implement a multi-pass solution. We first warp just the hair pixels with the hair cage (see Figs. 3 and 5). A flood-filled binary mask of the hair outline supports the query $isHair(x,y)$ in the source image. We then fill the rest of the pixels, opting for multi-cage blending for non-component (e.g. skin) pixels, and single cage warping for component pixels (e.g. eyes, nose, mouth, ears). We do not use multi-cage blending for the latter in order to maintain strong local shape influence. For multi-cage influence, we convert a destination image pixel p into unique polar coordinates $\{s,t\}_i$ for $i \in I$, where I is the set of cages in pixel p 's neighborhood. The blending weights are derived as inversely proportional to pixel p 's distance to the nearest boundary point on the associated cage, as shown in Eqn. 2, where d_i is the distance between p and the center of $cage_i$, and r_i is the corresponding center-boundary distance (see Sec. 3).

$$b_i = \frac{1}{d_i - r_i}. \quad (2)$$

The consequence of the above distance function is that the upper face and jaw cages dominate more towards the silhouette, while the face component cages (e.g. eyes, brows, nose and mouth) dominate in the center. Minor errors occurring due to skin pixels landing up in component regions, are easily corrected by fetching the nearest boundary pixel (an O(1) lookup from the cage's $\{t,r\}$ cache). The multi-pass algorithm above is able to generate a wide variety of warps with negligible artifacts.

7 Conclusions and Future Work

We have successfully demonstrated three different strategies for using primitive shapes to generate face caricatures: 1) Golden Ratio exaggeration; 2) Art rules; 3) Constrained adaptation to cartoon templates. The mesh-less parametric image warp algorithm produces decent results.

We have shown that our results (using synthesis method 1) produce more plausible results than an existing learning based approach [19]. We have also successfully stylized a neutral expression Caucasian man with Mean, Devious and Innocent caricature rules. Lastly, we have adapted a Caucasian and Asian man to single cartoon templates drawn by our in-house artist. This is a viable alternative to learning patterns from large collections of artwork, which could still produce unidentifiable cartoons. We have added an additional shape dimension to the caricature problem, decoupled from Golden Ratio Mean distances. We are able to detect the “amount” of primitive-shapedness in features, and are hence able to amplify this amount. There is also a lot

of scope for adding more personality based shape operations to create expressive shape-psychology based caricatures. We are currently working on incorporating layman perception feedback into the cartoon generation loop. We also hope to include Golden Ratio [20] for different races. Lastly, we are exploring hair-shape stylization as an additional feature that could enhance cartoon personality.

Acknowledgement

We would like to thank Kechit Goyal and Naval Chopra, Computer Science sophomores from IIT Delhi and IIT Bombay respectively, for their early work on face shape stylization. Paraag Bhatnagar, a secondary student in Singapore, drew the template cartoons in Fig. 8. This research is part of a Media Development Authority of Singapore (MDA GAMBIT) funded project, titled *Computational Aesthetics: Shape Psychology in Character Design*, WBS: R252000357490.

References

- [1] Akleman, E.: Making caricature with morphing. In: ACM SIGGRAPH 1997, p. 145 (1997)
- [2] Ashraf, G., Nahiduzzaman, K.M., Le, N.K.H., Mo, L.: Drafting 2D Characters with Primitive Shape Scaffolds. In: Second International Conference on Creative Content Technologies, Computation World, Lisbon (November 2010) (in press)
- [3] Bancroft, T.: Creating Characters with Personality. Watson-Guptill (2006)
- [4] Beiman, N.: Prepare to Board! Creating Story and Characters for Animated feature. Focal Press (2007)
- [5] Brennan, S.: Caricature generator., Master's thesis, Cambridge, MIT (1982)
- [6] Chen, H., Xu, Y., Shum, H., Zhu, S., Zheng, N.: Example based facial sketch generation with non-parametric sampling. In: ICCV 2001, pp. 433–II 438 (2001)
- [7] Chiang, P.-Y., Liao, W.-H., Li, T.-Y.: Automatic Caricature Generation by Analyzing Facial Features. In: Asian Conference on Computer Vision, Korea, January 27-30 (2004)
- [8] Criminisi, A., Perez, P., Toyama, K.: Object Removal by Exemplar-based Inpainting”, Madison. In: WI Proc. IEEE Computer Vision and Pattern Recognition (June 2003)
- [9] Freeman, W.T., Tenenbaum, J.B., Pasztor, E.: An example-based approach to style translation for line drawings, Technical Report 11, MERL Technical Report, Cambridge, MA (February 1999)
- [10] Garrett, L.: Visual Design: A Problem Solving Approach. RE Krieger Pub.Co., Huntington (1975)
- [11] Gooch, B., Reinhard, E., Gooch, A.: Human facial illustrations: Creation and psychophysical evaluation. ACM Trans. Graph. 23(1), 27–44 (2004)
- [12] Hart, C.: Cartoon Cool: How to Draw New Retro Style Characters. Watson-Guptill (2005)
- [13] Igarashi, T., Moscovich, T., Hughes, J.F.: As-rigid-as possible shape manipulation. ACM Trans. Graphics 24(3), 1134–1141 (2005)
- [14] Islam, M.T., Nahiduzzaman, K.M., Why, Y.P., Ashraf, G.: Learning from Humanoid Cartoon Designs. In: Perner, P. (ed.) ICDM 2010. LNCS, vol. 6171, pp. 606–616. Springer, Heidelberg (2010)

- [15] Iwashita, S., Takeda, Y., Onisawa, T.: Expressive facial caricature drawing. In: IEEE International Conference on Fuzzy Systems, vol. 3, pp. 1597–1602 (1999)
- [16] Li, Y., Kobatake, H.: Extraction of facial sketch based on morphological processing. In: IEEE International Conference on Image Processing, vol. 3, pp. 316–319 (1997)
- [17] Librande, S.E.: Example-based character drawing., Master's thesis, Cambridge. MA. MIT (1992)
- [18] Liang, L., Chen, H., Xu, Y.-Q., Shum, H.-Y.: Example-based Caricature Generation with Exaggeration. In: IEEE Proceedings of the 10th Pacific Conference on Computer Graphics and Applications (2002)
- [19] Liu, J., Chen, Y., Gao, W.: Mapping Learning in Eigenspace for Harmonious Caricature Generation. In: ACM Multimedia, pp. 683–686 (2006)
- [20] Marquardt, S.R.: Marquardt Beauty Analysis. MAI US patent no. 5.867588
- [21] McCloud, S.: Understanding Comics: The Invisible Art. Kitchen Sink Press, Princeton (1993)
- [22] Nishino, J., Kamyama, T., Shira, H., Odaka, T., Ogura, H.: Linguistic knowledge acquisition system on facial caricature drawing system. In: IEEE International Conference on Fuzzy Systems, vol. 3, pp. 1591–1596 (1999)
- [23] Web Ref, Aesthetics in international beauty pageants,
<http://www.femininebeauty.info>
- [24] Web Ref, <http://www.myheritage.com>
- [25] Weng, Y., Xu, W., Wu, Y., Zhou, K., Guo, B.: 2D shape deformation using nonlinear least squares optimization. The Visual Computer 22(9), 653–660 (2006)

i-m-Breath: The Effect of Multimedia Biofeedback on Learning Abdominal Breath

Meng-Chieh Yu¹, Jin-Shing Chen³, King-Jen Chang⁴, Su-Chu Hsu⁵,
Ming-Sui Lee^{1,2}, and Yi-Ping Hung^{1,2}

¹ Graduate Institute of Networking and Multimedia,
National Taiwan University, Taipei, Taiwan

² Department of Computer Science & Information Engineering,
National Taiwan University, Taipei, Taiwan

³ College of Medicine,
National Taiwan University, Taipei, Taiwan

⁴ Department of Surgery,
Cheng Ching General Hospital

⁵ Department of New Media Art,
Taipei National University of the Arts, Taiwan

Abstract. Breathing is a natural and important exercise for human beings, and the right breath method can make people healthier and even happier. *i-m-Breath* was developed to assist users in learning of abdominal breath, which used Respiration Girth Sensors (RGS) to measure user's breath pattern and provided visual feedback to assist in learning abdominal breath. In this paper, we tried to study the effect of biofeedback mechanism on learning of abdominal breath. We cooperated with College of Medicine in National Taiwan University to take the experiments to explore whether the biofeedback mechanism affect the learning of abdominal breath. The results of the experiments showed that *i-m-Breath* could help people in improving the breath habit from chest breath to abdominal breath, and in the future the system will be used the hospital. Finally, this study is important for providing a biofeedback mechanism to assist users in better understanding of his breath pattern and improving the breath habit.

Keywords: Abdominal Breath, Multimedia Biofeedback, Optoelectronic Plethysmography (OEP).

1 Introduction

Breathing is a natural behavior, which occurs without conscious control. Therefore, the breathing behaviors are very important for the maintenance of our physical health and the balance of various physiological functions. When the breathing depth decrease and can't provide enough oxygen and energy, it disables the body to eliminate its fatigue and makes us nervous and upset. Many studies show that breathing is a fundamental behavioral manifestation of the psychological and physiological state of human beings [2]. Blumenstein [7] also demonstrated that breath pattern is strongly associated with techniques for the regulation of mental states. Besides, many studies

also show that appropriate breath habit can reduce the times of asthmatic attack [9], delay the deterioration of chronic obstruction of pulmonary diseases [10], and reduce the probability of post-operative pain and complications on patients [11]. Abdominal breath, commonly known as diaphragmatic breathing, is a recommended breath method. Abdominal breath is very beneficial for the body, because breath abdominally allows a good quantity of oxygen in to your lungs. Therefore, this study aimed to develop a system that can assist users in learning of abdominal breath.

In the part of breath detection, there are many techniques in detecting the breath pattern and analyzing the breath information, such as Optoelectronic Plethysmography (OEP) [13], Ultra Wideband (UWB) [14], Spirometer, Gas Analyzer, Respiration Girth Sensor (RGS), and ECG-derived respiration technique [15]. Traditional training of breath learning demands one-by-one instruction and lacks of standardized evaluation. In this paper, we describe the breath-aware garment and biofeedback mechanisms of *i-m-Breath*, and use the training of abdominal breath as an example application to evaluate the effectiveness of it. The main purpose of this study is to determine if the use of biofeedback mechanisms would improve inappropriate breath habit of the users, and toward to recommend breathe habit – abdominal breath. In view of the preceding research purpose, two major sets of research questions to be addressed in this study are as follows: (a) whether *i-m-Breath* can assist users in learning abdominal breath, and (b) whether *i-m-Breath* can improve the breath habit and increase the total volume of breath. The goal of this study is to develop and verify an effective biofeedback mechanism to assist people to breathe correctly, and toward to the ideal of preventive medicine. In this paper, we have organized the rest of this paper in the following way: the first section of the article is a review of the literature, addressing both breath detection methods and the application in breath learning. This is followed by the introduction of different measurement of breath pattern. The third section describes the methodology and procedures for the collection of breathing data on multimedia biofeedback. The results for the various analyses are presented following each of these descriptive sections. Finally, conclusions are presented and suggestions are made for further research.

2 Related Work

A traditional method of breathing rehabilitation and breath learning needs professional respiratory therapist for learning assistance. The treatment requires lots of manpower and lack of quantitative evaluation of the progress during the period of breath learning. However, with the development of the information technology and human-computer interaction in recent years, some studies used multimedia and biofeedback mechanisms to assist users in the field of rehabilitation, self-healing, and health-care. Wild Divine Company [16] developed a well known self-healing system which used interactive animations and sounds to help users to control the breath and get better conditions both in mind and body. Morris's research [12] aimed to help people tune in to early signs of stress and modulate reactivity that could potentially damage their relationships and long-term health, and developed mobile therapies which provided just-in-time coaching. Thought Technology Ltd. [3] developed a biofeedback system which detected user's vital signal and visualized to the form of interactive multimedia. Besides, Venkat [8] used accelerometer which was put on user's abdomen to detect the movement of abdomen, and displayed the breath information. Kaushik [4] proved that

biofeedback assisted abdominal breath and systematic relaxation was very useful in migraine and had significantly better long-term prophylactic effect. Pastor [5] obtained demonstrated that biofeedback combined with this breath pattern produced a significant reduction in psycho physiological activation and improved learning through biofeedback techniques. However, none of related research in our survey has a complete study to prove the effect of the proposed biofeedback mechanisms, and whether the user has actually improved breath habits. In this study, a complete research on integrating breath-aware garment and multimedia feedback for improving breath habit was proposed. In this paper, we will describe the measurement of breath signal, the breath-aware garment and the calibration mechanisms of sensing signal, the biofeedback mechanisms, the experimental design, and the experimental results.

3 System Framework

i-m-Breath includes a breath-aware garment to detect user's breath pattern, an wireless transition module by bluetooth, and a desktop system to analysis the breath signal and provide visual feedback. In this section, the method of breath detection and biofeedback mechanisms of *i-m-Breath* will be introduced.

3.1 Breath Detection

There are many techniques in detecting the breath signal. This study used the technique of Respiration Girth Sensors (RGS) [21] and OEP [13] to detect participants' breath pattern. The method of RGS is used in *i-m-Breath*, and it has the advantage that it is portable and easy to use. The method of OEP is used as the evaluation tool in this study. OEP has the advantage that it can measure respiratory information accurately, and can measure breath volume in chest and abdomen separately. The following will introduce the detection method and the calibration procedure separately.

In the part of RGS, *i-m-Breath* uses Procomp Infinity [21] and two respiration girth sensors (RGS) to detect user's chest and abdominal elongation while breath. The sample rate is 25 samples per second. Before breath detection, the RGS sensor would be calibrated to translate the detection unit from raw data of ProComp2 to unit of length (cm). In each RGS, a mapping table was established to record the relationship between the value of the raw data and the unit of length. The default length of RGS is about 6 cm in unforced condition, and the pre-tensor of RGS is about 12 cm. Therefore, a ruler is used to calibrate the RGS from 12 cm to 25 cm. The system records the raw data with increment of RGS in 0.5 cm which pulls by hand (see Fig. 1a). Finally, the mapping table will have 27 paired mapping values to translate the raw data to the unit of length. Fig. 1b shows the mapping table of one RGS. The RGS is a sensitive *girth sensor* using a latex rubber band, and it could have the problem of fatigue. For this reason, the RGS could calibrate periodically. In this study, the RGS was calibrated every day during the experiment.

After calibrating the measurement unit of RGS, users should wear the breath-aware garment and adjust the wearing tightness and breathe parameters. In order to keep users not to wear the RGS too tight or too loose, the minimal breath elongation should in the period of 12~13 cm. In the calibration procedure, users are asked to exhale

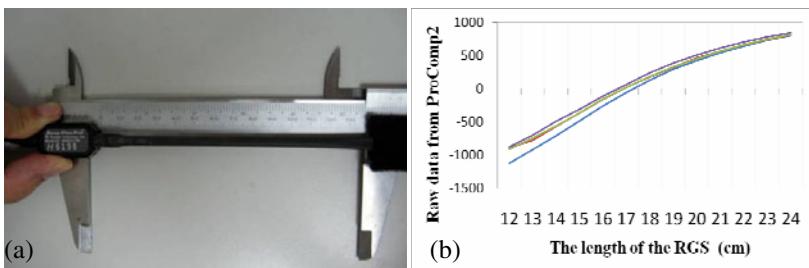


Fig. 1. Calibration of the RGS from raw data to the unit of length. (a) Using a ruler to calibrate the RGS every 0.5 cm; (b) The mapping table of the RGS four times.

extremely to get the minimal breath elongation. If the length is less than 12 cm, users should tighten the RGS, and if the length is greater than 13 cm, users should loosen the RGS. The procedure of wearing tightness would repeat until the minimal breath elongation is between 12~13 cm.

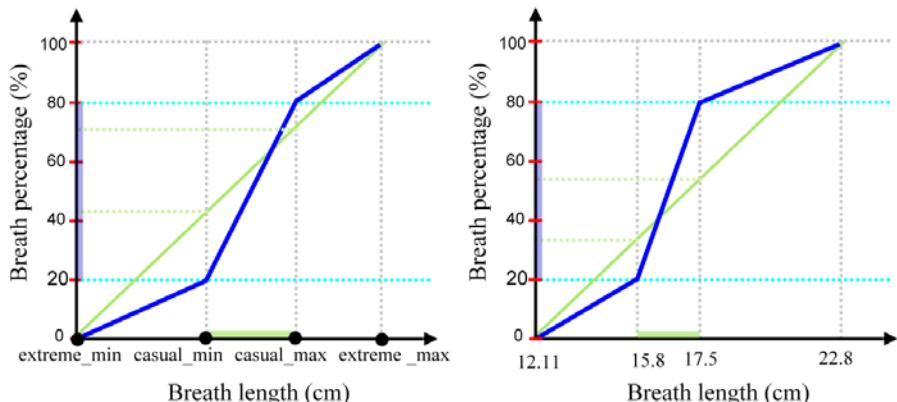


Fig. 2. Calibration of the breath length and the dynamic range. Green line indicates the result after linear interpolation. Blue line indicates the result after the calibration of dynamic range. (a) The diagram of the calibration; (b) An actual calibration result and the dynamic range by a participant.

Besides, because the dynamic range of user's breath elongation is small in general condition, and even he could not use abdominal breath, it could ineffective in the mechanism of biofeedback. Therefore, this system adopts a procedure to calibrate the dynamic range of user's breathe. In this calibration procedure, users are asked to calibrate the length of RGS in the condition of maximal breath extremely, minimal breath casually, and maximal breath casually separately. We set the range of breath from minimal breath extremely to maximal breath extremely. The dynamic range of casual breath (minimal breath casually ~ and maximal breath casually) is 60% (20%~80) of the length of deep breath, and others has 40% (0~20% and 80~100%). An experiment was adopted and the results showed that the calibration procedure of dynamic range

could enhance the effect 2.2 times than the method of linear interpolation. After the procedure, users could interact with the system more easy and effective. Fig. 2 shows the calibration chart of breath length and dynamic range.

Another technique of breath detection used in this study is OEP [13] (see Fig. 3). In the experiment, in order to detect user's breathing pattern accurately, OEP was selected as the evaluation tool. OEP System can measure the volume of user's chest wall and its variation during breathing, using reflective and non invasive markers attached to the thoraco-abdominal skin by biadhesive hypoallergenic tape. The three-dimensional positions of the markers are obtained thanks to infrared light video cameras with flashing LED's. OEP System measures the different compartments of the chest wall, provides the continuous monitoring of all ventilatory parameters. In our experiment, OEP System was used to measure participants' breath pattern, including chest volume, abdominal volume.



Fig. 3. Breath detection by using OEP equipment. (a) A participant is testing the breath pattern; (b) Detecting the breath pattern and visualized in real-time.

3.2 Biofeedback

While user already wear the breath-aware garment and calibrate the wearing tightness of abdominal RGS and chest RGS, he can start to use the biofeedback system to learn abdominal breath. This study adopted two phases of breath learning, including a real-time reflection mechanism to help user in learning abdominal breath, and a real-time guidance mechanism to assist user in following the regularity of suitable breath pattern. Besides, a ratio chart was representative of the chest/abdomen ratio with user's breath pattern while breath. The following would introduce the mechanisms of *i-m-Breath*. In the mechanism of real-time reflection, an animation of a virtual frog was representative of the abdominal breath situation with user's inhalation and exhalation. The animation of virtual frog was designed by a 3D model, and in order to decrease the computation of the system, we rendered 100 static pictures with a sequential movement of the frog. The resolution in each picture was 1024 pixels in width and 768 pixels in depth, and the sampling rate of the system was 15 frames per second. When inhaling, the virtual frog would plump its belly; when exhaling, the virtual frog would shrink its belly (see Fig. 4).

Respiratory pacing is an easily learned self-control strategy and potentially may be a useful therapeutic tool [6]. In the mechanism of real-time guidance, a breathing curve was representative of the guidance line and a light spot would fly on the path of

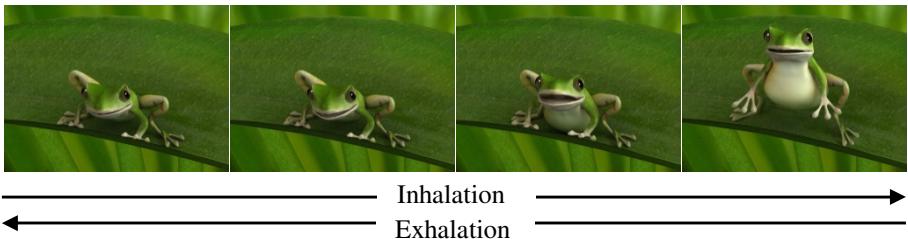


Fig. 4. The reflection mechanism to learn abdominal breath: An animation of a virtual frog to represent user's breathe condition on abdomen

the curve, which could guide user to breathe with its rhythm. The light spot was the guidance of specified breathing frequency. The light spot will move with a dotted curve which shows the guidance of the breath pattern. Besides, a virtual lady bug was representative of user's status of abdominal breath. While user inhaling, the virtual lady bug would fly higher, and while user exhaling, the virtual lady bug would fly lower. The rhythm of breathing curve can be changed by users, doctors or respiratory therapists to set a appropriate guidance of breath frequency (see Fig. 5). Moreover, *i-m-Breath* provided an evaluation mechanism. While user's breath followed with the curve of breathing guidance well, and the flower would bloom. On the contrary, while user did not follow with the guidance of breath curve and the flower would wither.

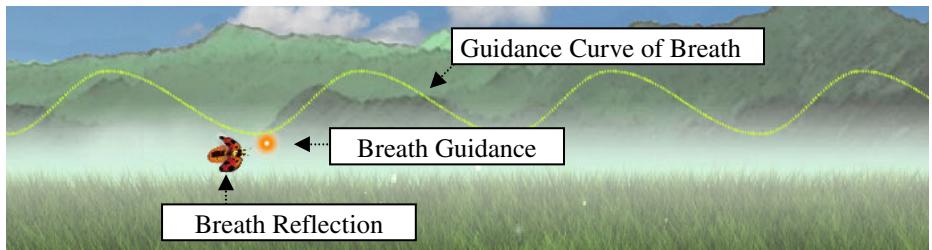


Fig. 5. The guidance mechanism to train the regularity of breath: The yellow curve shows the path of the breath pattern; the movement of light spot shows the guidance of breath depth; the movement of lady bug reflects user's abdominal breath depth

Therefore, *i-m-Breath* also provided a diagraph both in the above two phases to show the chest/abdomen chart to assist users in understanding of his breath method. The diagraph could display the relationship between user's chest breathing and abdominal breath. The X-axis indicated the breath depth of user's abdominal breath, and Y-axis indicated the breath depth of user's chest breathing. The location of the yellow ball was composed by chest breath depth in X-axis and abdominal breath depth in Y-axis in the same time (see Fig. 6). If user uses abdominal breath more than chest breath, the yellow ball would locate at lower right of the chart (Fig. 6a). On the contrast, if user uses chest breath more than abdominal breath, and the yellow ball will locate at higher left of the chart (Fig. 6b). However, diagraph could also show the information of breath depth. If user uses shallow breath , and the yellow ball will locate at lower right of the chart (Fig. 6c). If user uses deep breath , and the yellow ball will locate at higher right of the chart (Fig. 6d). The breath ratio chart shows in the lower left corner of the screen both in frog animation and flower animation.

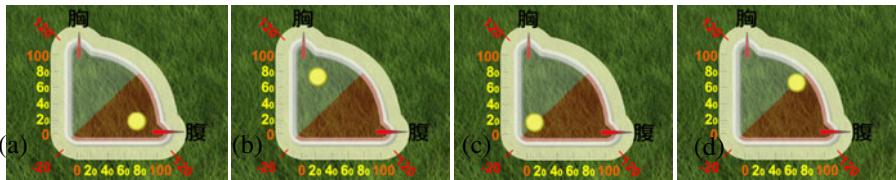


Fig. 6. The diagram of the relationship between chest breath and abdominal breath. (a) shows that user used abdominal breath; (b) shows that user used chest breath; (c) shows that user used shallow breath; (d) shows that user used deep breath.

4 Experimental Methods

4.1 Participants and Location

The participants volunteered to participate in this experiment, and the average age of them is 24 ($SD=3.71$). All participants did not have the experience of learning abdominal breath. There were five participants in experiment group, and five participants in control group. During the experiment, we ensured that all participants did not get sick during experiment. Besides, the gender of all participants was male because they were asked to undress on upper body to adhesive twenty reflective balls on body while using OEP test. The experiment was held in two places, including a department of respiratory measurement at National Taiwan University hospital and a laboratory in the faculty building at National Taiwan University. The participants in experimental group were asked to learn the abdominal breath by using *i-m-Breath* at laboratory for one month, and the participants in control group were asked to learn the abdominal breath himself.

4.2 Experimental Procedure

In this experiment, all participants were asked to learn the method of abdominal breath for one month. In order to analysis the breath pattern accurately, participants were asked to measure the breath pattern by using OEP equipment in day one, day two, and day thirty-one in the National Taiwan University Hospital. In day one of the experiment, participants' original breath pattern was recorded. In day two, a senior respiratory therapist taught all participants the skill of abdominal breath, and then tested the breath pattern by using OEP equipment, too. In day three to day thirty-one, all participants were divided into two groups, experimental group and control group. Five participants in experimental group were asked to use *i-m-Breath* to practice abdominal breath for 20 minutes once every two days, and another days they were asked to practice abdominal breath themselves. Besides, five participants in control group were asked to learn abdominal breath every day, and twenty minutes in each time. Fig. 7 shows the experimental procedure in this study.

We reminded the participants in control group to practice abdominal breath every week and give them a teaching manual of abdominal breath to ensure that they had practiced abdominal breath on time. Finally, all participants were asked to test the breath pattern by using OEP equipment in day thirty-one. While testing, all participants were asked not to breathe deliberately. None of the participants were blind as to the nature of the experiment. They were not told, however, what types of results were expected.

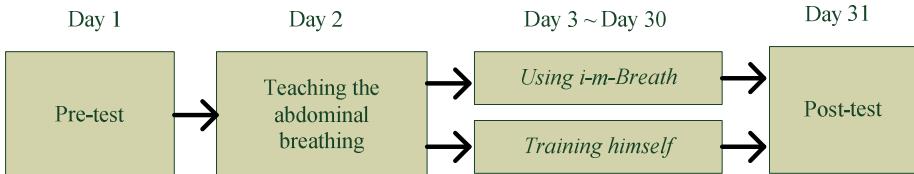


Fig. 7. Experimental procedure

4.3 Experimental Results

We analyzed participants' average volume of breath in chest and abdomen by using OEP equipment, and also analyzed their breath pattern (total volume of breath and chest-abdominal breath ratio) both in experimental group and control group. All participants were measured the breath pattern in day 1 (pre-test), day 2 (after teaching by respiratory therapist), and day 31 (post-test). They were asked to breathe naturally and not to breathe deliberately. Each measurement time was ten minutes by using OEP equipment.

In the comparison of abdominal breath, the average volume of abdominal breath was 45 ml per minute in experimental group, and 54 ml per minute in control group in the first day. The result showed that there were no significant different in abdominal breath between experimental group and control group in the beginning of the experiment (See Fig. 8a). In the second day, after teaching of abdominal breath by respiratory therapist, the average volume of abdominal breath was 60 ml per minute in the experimental group, and 63 ml per minute in the control group. The result showed that learning of abdominal breath by respiratory therapist had significant effects increasing the volume of abdominal breath both on the experimental group ($p < .005$) and the control group ($p < .05$). After one month, the average volume of abdominal breath was 80.5 ml per minute in the experimental group, and 54.2 ml per minute in the control group. The result showed that there was significant increase on the experimental group after one-month experiment ($p < .005$), but there was no significant effect on the control group. In the comparison of chest breath, the average volume of chest breathing was 38.3 ml per minute in the experimental group, and 30.6 ml per minute in the control group in the first day. The result showed that there were no significant different on abdominal breath between the experimental group and the control group in the beginning of the experiment (see Fig. 8b). In the second day, after teaching of abdominal breath by professional respiratory therapist, the average volume of abdominal breath was 24.8 ml per minute in experimental group, and 20.5 ml per minute in the control group. The result showed that learning of abdominal breath by respiratory therapist did not have significant effect both in the experimental group and the control group. After one month, the average volume of abdominal breath was 8.9 ml per minute in the experimental group, and 31 ml per minute in the control group. The result showed that there was significant decrease on the experimental group after one-month experiment ($p < .005$), but there was no significant effect on the control group.

In the comparison of total volume of breath, the average total volume of breath was 83 ml per minute in the experimental group, and 84 ml per minute in the control group in the first day. The result showed that there was no significant different

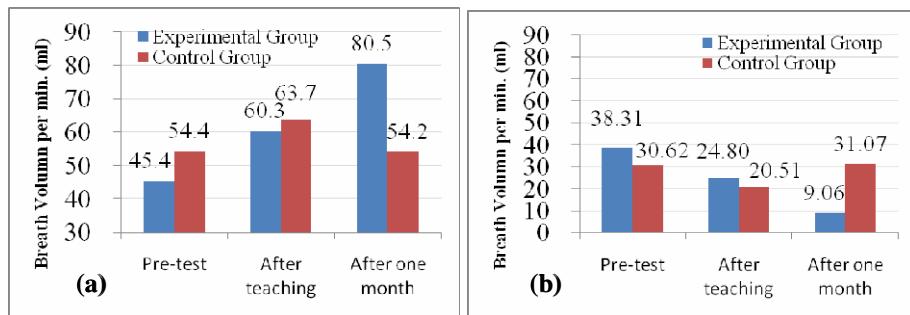


Fig. 8. The comparison of breath volumn between pre-test (day 1), after teaching, and after one month. (a) the comparison of abdominal breath volume per min; (b) the comparison of chest breath volume per min

between the experimental group and the control group in the beginning of the experiment (See Fig. 9a). In the second day, after the teaching of abdominal breath by respiratory therapist, the average volume of abdominal breath was 85 ml per minute in the experimental group, and 84 ml per minute in the control group. The result showed that learning of abdominal breath by respiratory therapist did not have significant effect on change the total volume of breath both in the experimental group and the control group. After one month, the average volume of abdominal breath was 89 ml per minute in the experimental group, and 85 ml per minute in the control group. Overall, the result showed that the total volume of breath increases in the experimental group and the control group, but there were no significant effects.

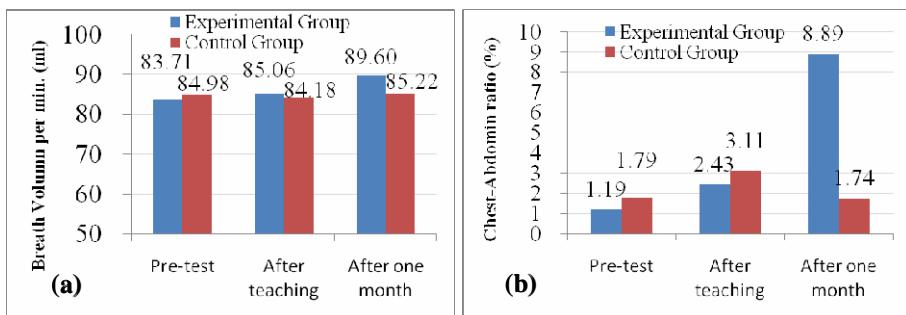


Fig. 9. The comparison of overall Breath pattern between pre-test, after teaching, and after one month. (a) the comparison of total breath volume per minute; (b) the comparison of chest-abdominal breath

The ratio of chest-abdominal breath is defined as volume of abdominal breath divided by volume of chest breath. The more the value means the more the abdominal breath be used. In the comparison of chest-abdominal breath ratio, the average chest-abdominal breath ratio was 1.19 in experimental group, and 1.79 in control group in the first day. The result showed that there was no significant difference between the experimental group and the control group in the beginning of the experiment. In the second day, after the teaching of abdominal breath by respiratory therapist, the average

ratio of chest-abdominal breath was 2.43 in the experimental group, and 3.11 in the control group. After one month, the average ratio of chest-abdominal breath was 8.89 in the experimental group, and 1.74 in the control group (see Fig. 9b). The result showed that the chest-abdominal breath ratio had significantly increased in the experimental group after the one-month experiment ($p<0.005$), but there was no significant effect on the control group.

In conclusion, the results of the abdominal breath volume, total volume of breath, and the ratio of abdomen-chest breath in this study showed that *i-m-Breath* had significant effect on the learning of abdominal breath. According to the observation of the experiment, participants could not control the frog easily in the beginning. However, they could control the frog easily after several days. Besides, the mechanism of breathing guidance is a complex problem, because there are many factors that could affect user's breath pattern, such as user's body type, exercise condition, emotion condition, etc. If the system provides improper guidance of breath pattern, users might get dizzy and uncomfortable. Therefore, we asked participants to set the time of breathing guidance, including the inhaling time and exhaling time.

6 Conclusions and Future Work

In this paper, we proposed *i-m-Breath* system with multimedia biofeedback and we have cooperated with College of Medicine in National Taiwan University to take the experiments to explore whether the biofeedback mechanism affect the learning of abdominal breath. The results of the experiments showed that *i-m-Breath* could help people in improving the breath habit from chest breath to abdominal breath. We had the following conclusions: (a) in our experiments, it showed the obvious and immediate effect to improve abdominal breath during learning with the advise of respiratory therapist around, (b) our biofeedback mechanism could increase users' interest and fun for helping the long-term breath learning and improve the breath habit effectively, and it is our main contribution which will be used the hospital. In our research, the participants were all healthy young man, and this study does not prove that the biofeedback mechanism has effect on other specific-groups, too. Besides, our study did not test the effect of biofeedback on the regularity of breath pattern. These problems will be improved in the future. Currently, we have developed a portable biofeedback system to assist users in breath learning anywhere they want. For the future research, we will develop a high accuracy of breath detection method and investigate the relationship between activity and breathing more precisely. In addition, we will integrate *i-m-Breath* system, activity recognition system, and multimedia biofeedback to develop a health care system in daily life.

Acknowledgment

This work was supported in part by the National Science Council, Taiwan, under grants NSC 98-2221-E-002-127-MY3, and by the Technology Development Program for Academia, Ministry of Economic Affairs, Taiwan, under the grant 98-EC-17-A-19-S2-0133. And thanks to Samwell Testing INC. who supported the OEP system.

References

1. Siepmann, M., Aykac, V., Unterdörfer, J., et al.: A pilot study on the effect of heart rate variability biofeedback in patients with depression and in healthy subjects. *Appl. Psychophysiol. Biofeedback* 33(4), 195–201 (2008)
2. Ley, R.: An introduction to the psychophysiology of breathing. *Applied Psychophysiology and Biofeedback* 19(2) (1994)
3. Thought Technology Ltd, <http://www.thoughttechnology.com>
4. Kaushik, R., Kaushik, R.M., Mahajan, S.K., Rajesh, V.: Biofeedback assisted diaphragmatic breathing and systematic relaxation versus propranolol in long term prophylaxis of migraine. *Complementary Therapies in Medicine* 13, 165–174 (2005)
5. Pastor, M.C., Menéndez, F.J., Sanz, M.T., Abad, E.V.: The Influence of Respiration on Biofeedback Techniques. *Appl. Psychophysiol. Biofeedback* 33, 49–54 (2008)
6. Clark, M.E., Hirschman, R.: Effect of paced respiration on anxiety reduction in a clinical population. *Biofeedback and Self-Regulation* 15(3), 273–285 (1990)
7. Blumenstein, B., Breslav, I., Bar-Eli, M., Tenenbaum, G., Weinstein, Y.: Regulation of Mental States and Biofeedback Techniques: Effect of Breathing Pattern. *Biofeedback and Self-Regulation* 20, 169–183 (1995)
8. Venkat, R.B., Sawant, A., Suh, Y., George, R., Keall, P.J.: Development and preliminary evaluation of a prototype audiovisual biofeedback device incorporating a patient-specific guiding waveform. *Phys. Med. Biol.* 53, 197–200 (2008)
9. Slader, C.A., Reddel, H.K., Spencer, L.M.: Double blind randomized controlled trial of two different breathing techniques in the management of asthma. *Thorax* 61, 651–656 (2006)
10. Guell, R., Resqueti, V., Sangenis, M.: Impact of pulmonary rehabilitation on psychosocial morbidity in patients with severe COPD. *Chest* 129, 899–904 (2006)
11. Westerdahl, E., Lindmark, B., Eriksson, T., Friberg, O., Hedenstierna, G., Tenling, A.: Deep-breathing exercises reduce atelectasis and improve pulmonary function after coronary artery bypass surgery. *Chest* 128, 3482–3488 (2005)
12. Morris, D., Brush, A.J.B., Brian, R.: Meyers SuperBreak: Using Interactivity to Enhance Ergonomic Typing Breaks. In: Proc. CHI 2008, pp. 1817–1826. ACM Press, New York (2008)
13. Aliverti, A., Dellacà, R.L., Pelosi, R., Chiumello, D., Pedotti, A., Gattinoni, L.: Optoelectronic plethysmography in intensive care patients. *Am. J. Resp. Crit. Care Med.* 161, 1546–1552 (2000)
14. Ossberger, G., Buchegger, T., Schimback, E., Stelzer, A., Weigel, R.: Non-invasive respiratory movement detection and monitoring of hidden humans using ultra wideband pulse radar. In: Proc. of the 2004 International Workshop on Ultra Wideband Systems Joint with Conference on Ultra Wideband Systems and Technology, pp. 395–399 (2004)
15. Moody, G.B., Mark, R.G., Zoccola, A., Mantero, S.: Derivation of respiratory signals from multi-lead ECGs. *Computers in Cardiology* 12, 113–116 (1985)
16. WildDivine Company, <http://www.wilddivine.com>

Author Index

- Ai, Mingjing II-383
Ardabilian, Mohsen I-206
Ariki, Yasuo II-454
Ashraf, Golam I-536
Assent, Ira I-140
Azad, Salahuddin II-442
Bailer, Werner I-359, II-219
Beecks, Christian I-140, I-381
Benoit, Alexandre I-350
Boeszoermenyi, Laszlo I-129
Boll, Susanne I-84
Cai, Junjie II-77
Cai, Qiyun II-393
Chan, Antoni B. I-317
Chang, King-Jen I-548
Chang, Kuang-I II-432
Chang, Richard I-328
Chang, Shih-Ming I-514, II-296
Chao, Hui II-65
Chen, Bing-Yu I-73, I-435
Chen, Hsin-Hui II-168, II-252
Chen, Hua-Tsung II-315
Chen, Jie II-12
Chen, Jin-Shing I-548
Chen, Jyun-Long II-432
Chen, Kuan-Wen I-171
Chen, Liming I-206
Cheng, Hsu-Yung II-187
Cheng, Jian I-307
Cheng, Kai-Yin I-73
Cheng, Sheng-Yi I-457
Chu, Wei-Ta I-229
Chu, Xiqing II-359
Chua, Tat-Seng I-262, I-392
Chua, TeckWee I-328
Chung, Sheng-Luen II-177
Ciobotaru, Madalina I-350
Cock, Jan De I-29
Collmosse, John I-118
Cooray, Saman H. I-424
Dai, Feng I-21, I-51
Dai, Wang II-393
De Bruyne, Sarah I-29, II-486
Dejrabia, Chabane I-251
Ding, Jian-Jiun II-168, II-252
Doman, Keisuke II-135
Dong, Shichao II-476
Du, Ruo II-35
Duan, Ling-Yu II-12
El Sayad, Ismail I-251
Emmanuel, Sabu II-285
Fan, Chih-Peng I-10
Fan, Jianping II-46, II-111
Fan, Jingwen II-359
Fan, Songchun II-483
Fan, Xin II-483
Fang, Yuchun II-393
Fang, Yuming I-370
Feng, Xiaoyi II-46, II-111
Foley, Colum II-337
Fu, Chuan II-274
Fu, Qiang II-57
Gao, Sheng II-99
Gao, Wen II-12
Goto, Satoshi I-161
Gu, Hui-Zhen II-315
Gu, Zhouye I-40
Guo, Aiyuan II-421
Guo, Jing-Ming II-177
Guo, Jinlin II-337
Gurrin, Cathal II-337
Han, Tony X. II-1
Hang, Kaiyu I-413
He, Xiangjian II-35
He, Xun I-161
He, Ying I-217, II-371
Hoeffernig, Martin II-263
Hoi, Steven C.H. I-217, II-371
Hollemeersch, Charles-Frederik I-29, II-486
Hong, Wei-Tyng II-187
Hou, ZuJun I-328

- Hsiao, Pei-Yung II-208
 Hsu, Chiou-Ting I-503
 Hsu, Hui-Huang I-514, II-296
 Hsu, Su-Chu I-548
 Hsu, Winston H. II-146
 Hsu, Yu-Ming II-146
 Hua, Xian-Sheng I-107
 Huang, Chun-Kai I-435
 Huang, Di I-206
 Huang, Jingjing I-403
 Huang, Jiun-De II-168
 Huang, Shih-Ming II-326
 Huang, Shih-Shinh II-208
 Huang, Szu-Hao I-151
 Huang, Yea-Shuan I-457, I-525
 Hung, Chia-Jung I-435
 Hung, Shang-Chih I-151
 Hung, Yi-Ping I-171, I-548
 Hürst, Wolfgang II-157, II-230
- Ide, Ichiro II-135
 Ikenaga, Takeshi I-492
 Ionescu, Bogdan I-350
- Jabbar, Khalid I-182
 Jang, Lih-Guong I-446
 Jeng, Bor-Shenn II-187
 Ji, Rongrong II-12
 Jia, Wenjing II-35
 Jian, Er-Liang II-208
 Jiang, Xinghao II-359
 Jin, Jesse S. II-25
 Jin, Xiaocong I-492
 Jin, Xin I-161
 Jose, Joemon I-118
- Kaiser, Rene II-263
 Kankanhalli, Mohan II-285
 Katayama, Norio I-284
 Kuai, Cheng Ying II-135
 Kuo, Tzu-Hao I-73
- Lai, Shang-Hong I-151
 Lambert, Patrick I-350
 Lambert, Peter I-29, II-486
 Lao, Songyang II-337
 Lau, Chiew Tong I-40, I-370
 Lay, Jose I-296
- Le, Nguyen Kim Hai I-536
 Lee, Bu-sung I-40, I-370
 Lee, Chien-Cheng II-187
 Lee, Felix II-219
 Lee, Hyowon I-424
 Lee, Ming-Sui I-548
 Lee, Pei-Jyun I-171
 Lee, Su-Ling II-196
 Lee, Suh-Yin II-315
 Lee, Tung-Ying I-151
 Lee, Tzu-Heng II-168
 Leman, Karianto I-328
 Levy, David I-296
 Li, Bing II-12
 Li, Chu-Yung I-525
 Li, Fan I-470
 Li, Haizhou II-99
 Li, Haojie II-476
 Li, Meng-Luen I-229
 Li, Peng I-307
 Li, Tom L.H. I-317
 Li, Yangxi II-479
 Li, Yiqun I-262, II-421
 Li, Zechao I-307
 Liang, Chao II-88
 Liao, Zhuhua II-274
 Lien, Cheng-Chang I-446
 Lim, Joo Hwee II-421
 Lin, Che-Hung II-177
 Lin, Chia-Wen I-370
 Lin, Chin-Lin II-401
 Lin, Daw-Tung II-401
 Lin, Pao-Yen II-168, II-252
 Lin, Shou-De I-339
 Lin, Shouxun I-21
 Lin, Tsung-Yu I-151
 Lin, Weisi I-40, I-370
 Lin, Yen-Liang II-146
 Lin, Yu-Shin II-296
 Liu, Guizhong I-470
 Liu, Haiming II-241
 Liu, Jingjing I-481
 Liu, Siying II-421
 Liu, Yan I-1
 Liu, Yang II-88
 Lo, Hung-Yi I-339
 Lou, Chengsheng II-393
 Lu, Hanqing I-307, I-481
 Luo, Jie II-393
 Luo, Jiebo I-403

- Luo, Suhuai II-25
 Luo, Yong II-479
- Mahapatra, Amogh I-273
 Mak, Mun-Thye II-411
 Mark Liao, Hong-Yuan I-503
 Martinet, Jean I-251
 Matsuoka, Yuta I-96
 Mayer, Harald II-263
 Mei, Tao I-107
 Miao, Chen-Hsien I-10
 Mo, Hiroshi I-284
 Mulholland, Paul II-241
 Murase, Hiroshi II-135
- Nakamura, Naoto II-348
 Nguyen, Duc Dung II-371
 Nijholt, Anton II-122
- O'Connor, Noel E. I-424
 Okada, Yoshihiro II-348
 Ong, Alex I-182
 Ouji, Karima I-206
- Park, Mira II-25
 Peng, Jinye II-46, II-111
 Peng, Yu II-25
 Pham, Nam Trung I-328
 Plass-Oude Bos, Danny II-122
 Poel, Mannes II-122
 Poppe, Chris I-29, II-486
 Preneel, Bart I-62
- Qian, Xueming I-413
 Quah, Chee Kwang I-182
 Quénot, Georges I-240
- Rabbath, Mohammad I-84
 Ren, Reede I-118
 Roy, Sujoy II-411, II-465
 Rüger, Stefan II-241
- Safadi, Bahjat I-240
 Sandhaus, Philipp I-84
 Satoh, Shin'ichi I-284
 Schoeffmann, Klaus I-129
 Seah, Hock Soon I-182
- Seidl, Thomas I-140, I-381
 Shao, Ling I-1
 She, Lanbo I-403
 Shen, Chuxiong II-359
 Shih, Shen-En I-193
 Shih, Timothy K. I-514, II-296
 Shirahama, Kimiaki I-96
 Sim, Terence II-465
 Smeaton, Alan F. II-337
 Snoek, Cees G.M. II-230
 Song, Dawei II-241
 Song, Wei II-442
 Spoel, Willem-Jan II-230
 Srivastava, Jaideep I-273
 Srivastava, Ruchir II-465
 Su, Han II-483
 Su, Jia I-492
 Su, Yu-Jen II-432
 Sun, Tanfeng II-359
 Sun, Weifeng II-476
- Tai, Shih-Chao II-304
 Takahashi, Tomokazu II-135
 Takano, Shigeru II-348
 Takiguchi, Tetsuya II-454
 Tam, King Yiu I-296
 Tan, Cheng II-483
 Tang, Feng II-57
 Tang, Nick C. I-503
 Tao, Dacheng II-479
 Tian, Qi II-77
 Tian, Yonghong I-273
 Tjondronegoro, Dian II-442
 Tomin, Mate II-230
 Tong, Yubing I-240
 Tretter, Dan II-65
 Tsai, Chang-Lung II-304
 Tsai, Chun-Yu I-73
 Tsai, Hsin-Ming II-208
 Tsai, Joseph I-514
 Tsai, Joseph C. II-296
 Tsai, Ming-Hsiu I-446
 Tsai, Mu-Yu II-432
 Tsai, Wei-Chin II-315
 Tsai, Wen-Hsiang I-193
 Tsai, Ya-Ting I-446
 Tseng, Chien-Cheng II-196
 Tu, Meng-Qui II-208
 Tyan, Hsiao-Rong I-503

- Uehara, Kuniaki I-96
Uren, Victoria II-241
Urruty, Thierry I-251
Uysal, Merih Seran I-381
- Wagner, Claudia II-263
Walle, Rik Van de I-29, II-486
Wan, Kong Wah II-411
Wan, Xin I-273
Wang, Brian II-146
Wang, Dayong I-217
Wang, Hsin-Min I-339
Wang, Jinqiao I-481, II-12
Wang, Lin II-35
Wang, Minghui I-161
Wang, Yan I-107
Wang, Yue I-328
Wang, Yunhong I-206
Wang, Zengfu II-77
Wei, Xin I-492
Weng, Li I-62
Wezel, Casper van II-157
Why, Yong Peng I-536
Wu, Bin II-359
Wu, Peng II-57
Wu, Pengcheng II-371
Wu, Qiang II-35
Wu, Shu-Min II-432
- Xu, Changsheng II-1, II-88
Xu, Chao II-479
- Yan, Chenggang I-51
Yan, Shuicheng II-1, II-465
Yan, Zhe I-413
Yang, Chunlei II-111
Yang, Jar-Ferr II-326
Yang, Jing II-274
Yeh, Wei-chang II-35
Yu, Chih-Chang II-187
Yu, Jen-Yu II-315
Yu, Like I-21
Yu, Meng-Chieh I-548
Yu, Nenghai I-403
- Zha, Zheng-Jun I-262, II-77
Zhang, Guoqing II-274
Zhang, Huanchen II-476
Zhang, Hui I-1
Zhang, Jinyu II-483
Zhang, Peng II-285
Zhang, Qing I-470
Zhang, Shanfeng II-359
Zhang, Tong II-65
Zhang, Xinming II-88
Zhang, Yongdong I-21, I-51
Zhang, Zheng I-182
Zhao, Lili II-383
Zhao, Yi-Liang I-392
Zheng, Yan-Tao I-262, I-392
Zheng, Yu II-454
Zhou, Ning II-46
Zhou, Xiangdong I-392
Zhu, Guangyu II-1
Zhuang, Liansheng I-403

Author Index

- Ai, Mingjing II-383
Ardabilian, Mohsen I-206
Ariki, Yasuo II-454
Ashraf, Golam I-536
Assent, Ira I-140
Azad, Salahuddin II-442
Bailer, Werner I-359, II-219
Beecks, Christian I-140, I-381
Benoit, Alexandre I-350
Boeszoeremenyi, Laszlo I-129
Boll, Susanne I-84
Cai, Junjie II-77
Cai, Qiyun II-393
Chan, Antoni B. I-317
Chang, King-Jen I-548
Chang, Kuang-I II-432
Chang, Richard I-328
Chang, Shih-Ming I-514, II-296
Chao, Hui II-65
Chen, Bing-Yu I-73, I-435
Chen, Hsin-Hui II-168, II-252
Chen, Hua-Tsung II-315
Chen, Jie II-12
Chen, Jin-Shing I-548
Chen, Jyun-Long II-432
Chen, Kuan-Wen I-171
Chen, Liming I-206
Cheng, Hsu-Yung II-187
Cheng, Jian I-307
Cheng, Kai-Yin I-73
Cheng, Sheng-Yi I-457
Chu, Wei-Ta I-229
Chu, Xiqing II-359
Chua, Tat-Seng I-262, I-392
Chua, TeckWee I-328
Chung, Sheng-Luen II-177
Ciobotaru, Madalina I-350
Cock, Jan De I-29
Collmosse, John I-118
Cooray, Saman H. I-424
Dai, Feng I-21, I-51
Dai, Wang II-393
De Bruyne, Sarah I-29, II-486
Dejrabia, Chabane I-251
Ding, Jian-Jiun II-168, II-252
Doman, Keisuke II-135
Dong, Shichao II-476
Du, Ruo II-35
Duan, Ling-Yu II-12
El Sayad, Ismail I-251
Emmanuel, Sabu II-285
Fan, Chih-Peng I-10
Fan, Jianping II-46, II-111
Fan, Jingwen II-359
Fan, Songchun II-483
Fan, Xin II-483
Fang, Yuchun II-393
Fang, Yuming I-370
Feng, Xiaoyi II-46, II-111
Foley, Colum II-337
Fu, Chuan II-274
Fu, Qiang II-57
Gao, Sheng II-99
Gao, Wen II-12
Goto, Satoshi I-161
Gu, Hui-Zhen II-315
Gu, Zhouye I-40
Guo, Aiyuan II-421
Guo, Jing-Ming II-177
Guo, Jinlin II-337
Gurrin, Cathal II-337
Han, Tony X. II-1
Hang, Kaiyu I-413
He, Xiangjian II-35
He, Xun I-161
He, Ying I-217, II-371
Hoeffernig, Martin II-263
Hoi, Steven C.H. I-217, II-371
Hollemeersch, Charles-Frederik I-29, II-486
Hong, Wei-Tyng II-187
Hou, ZuJun I-328

- Hsiao, Pei-Yung II-208
 Hsu, Chiou-Ting I-503
 Hsu, Hui-Huang I-514, II-296
 Hsu, Su-Chu I-548
 Hsu, Winston H. II-146
 Hsu, Yu-Ming II-146
 Hua, Xian-Sheng I-107
 Huang, Chun-Kai I-435
 Huang, Di I-206
 Huang, Jingjing I-403
 Huang, Jiun-De II-168
 Huang, Shih-Ming II-326
 Huang, Shih-Shinh II-208
 Huang, Szu-Hao I-151
 Huang, Yea-Shuan I-457, I-525
 Hung, Chia-Jung I-435
 Hung, Shang-Chih I-151
 Hung, Yi-Ping I-171, I-548
 Hürst, Wolfgang II-157, II-230
- Ide, Ichiro II-135
 Ikenaga, Takeshi I-492
 Ionescu, Bogdan I-350
- Jabbar, Khalid I-182
 Jang, Lih-Guong I-446
 Jeng, Bor-Shenn II-187
 Ji, Rongrong II-12
 Jia, Wenjing II-35
 Jian, Er-Liang II-208
 Jiang, Xinghao II-359
 Jin, Jesse S. II-25
 Jin, Xiaocong I-492
 Jin, Xin I-161
 Jose, Joemon I-118
- Kaiser, Rene II-263
 Kankanhalli, Mohan II-285
 Katayama, Norio I-284
 Kuai, Cheng Ying II-135
 Kuo, Tzu-Hao I-73
- Lai, Shang-Hong I-151
 Lambert, Patrick I-350
 Lambert, Peter I-29, II-486
 Lao, Songyang II-337
 Lau, Chiew Tong I-40, I-370
 Lay, Jose I-296
- Le, Nguyen Kim Hai I-536
 Lee, Bu-sung I-40, I-370
 Lee, Chien-Cheng II-187
 Lee, Felix II-219
 Lee, Hyowon I-424
 Lee, Ming-Sui I-548
 Lee, Pei-Jyun I-171
 Lee, Su-Ling II-196
 Lee, Suh-Yin II-315
 Lee, Tung-Ying I-151
 Lee, Tzu-Heng II-168
 Leman, Karianto I-328
 Levy, David I-296
 Li, Bing II-12
 Li, Chu-Yung I-525
 Li, Fan I-470
 Li, Haizhou II-99
 Li, Haojie II-476
 Li, Meng-Luen I-229
 Li, Peng I-307
 Li, Tom L.H. I-317
 Li, Yangxi II-479
 Li, Yiqun I-262, II-421
 Li, Zechao I-307
 Liang, Chao II-88
 Liao, Zhuhua II-274
 Lien, Cheng-Chang I-446
 Lim, Joo Hwee II-421
 Lin, Che-Hung II-177
 Lin, Chia-Wen I-370
 Lin, Chin-Lin II-401
 Lin, Daw-Tung II-401
 Lin, Pao-Yen II-168, II-252
 Lin, Shou-De I-339
 Lin, Shouxun I-21
 Lin, Tsung-Yu I-151
 Lin, Weisi I-40, I-370
 Lin, Yen-Liang II-146
 Lin, Yu-Shin II-296
 Liu, Guizhong I-470
 Liu, Haiming II-241
 Liu, Jingjing I-481
 Liu, Siying II-421
 Liu, Yan I-1
 Liu, Yang II-88
 Lo, Hung-Yi I-339
 Lou, Chengsheng II-393
 Lu, Hanqing I-307, I-481
 Luo, Jie II-393
 Luo, Jiebo I-403

- Luo, Suhuai II-25
 Luo, Yong II-479
- Mahapatra, Amogh I-273
 Mak, Mun-Thye II-411
 Mark Liao, Hong-Yuan I-503
 Martinet, Jean I-251
 Matsuoka, Yuta I-96
 Mayer, Harald II-263
 Mei, Tao I-107
 Miao, Chen-Hsien I-10
 Mo, Hiroshi I-284
 Mulholland, Paul II-241
 Murase, Hiroshi II-135
- Nakamura, Naoto II-348
 Nguyen, Duc Dung II-371
 Nijholt, Anton II-122
- O'Connor, Noel E. I-424
 Okada, Yoshihiro II-348
 Ong, Alex I-182
 Ouji, Karima I-206
- Park, Mira II-25
 Peng, Jinye II-46, II-111
 Peng, Yu II-25
 Pham, Nam Trung I-328
 Plass-Oude Bos, Danny II-122
 Poel, Mannes II-122
 Poppe, Chris I-29, II-486
 Preneel, Bart I-62
- Qian, Xueming I-413
 Quah, Chee Kwang I-182
 Quénot, Georges I-240
- Rabbath, Mohammad I-84
 Ren, Reede I-118
 Roy, Sujoy II-411, II-465
 Rüger, Stefan II-241
- Safadi, Bahjat I-240
 Sandhaus, Philipp I-84
 Satoh, Shin'ichi I-284
 Schoeffmann, Klaus I-129
 Seah, Hock Soon I-182
- Seidl, Thomas I-140, I-381
 Shao, Ling I-1
 She, Lanbo I-403
 Shen, Chuxiong II-359
 Shih, Shen-En I-193
 Shih, Timothy K. I-514, II-296
 Shirahama, Kimiaki I-96
 Sim, Terence II-465
 Smeaton, Alan F. II-337
 Snoek, Cees G.M. II-230
 Song, Dawei II-241
 Song, Wei II-442
 Spoel, Willem-Jan II-230
 Srivastava, Jaideep I-273
 Srivastava, Ruchir II-465
 Su, Han II-483
 Su, Jia I-492
 Su, Yu-Jen II-432
 Sun, Tanfeng II-359
 Sun, Weifeng II-476
- Tai, Shih-Chao II-304
 Takahashi, Tomokazu II-135
 Takano, Shigeru II-348
 Takiguchi, Tetsuya II-454
 Tam, King Yiu I-296
 Tan, Cheng II-483
 Tang, Feng II-57
 Tang, Nick C. I-503
 Tao, Dacheng II-479
 Tian, Qi II-77
 Tian, Yonghong I-273
 Tjondronegoro, Dian II-442
 Tomin, Mate II-230
 Tong, Yubing I-240
 Tretter, Dan II-65
 Tsai, Chang-Lung II-304
 Tsai, Chun-Yu I-73
 Tsai, Hsin-Ming II-208
 Tsai, Joseph I-514
 Tsai, Joseph C. II-296
 Tsai, Ming-Hsiu I-446
 Tsai, Mu-Yu II-432
 Tsai, Wei-Chin II-315
 Tsai, Wen-Hsiang I-193
 Tsai, Ya-Ting I-446
 Tseng, Chien-Cheng II-196
 Tu, Meng-Qui II-208
 Tyan, Hsiao-Rong I-503

- Uehara, Kuniaki I-96
Uren, Victoria II-241
Urruty, Thierry I-251
Uysal, Merih Seran I-381
- Wagner, Claudia II-263
Walle, Rik Van de I-29, II-486
Wan, Kong Wah II-411
Wan, Xin I-273
Wang, Brian II-146
Wang, Dayong I-217
Wang, Hsin-Min I-339
Wang, Jinqiao I-481, II-12
Wang, Lin II-35
Wang, Minghui I-161
Wang, Yan I-107
Wang, Yue I-328
Wang, Yunhong I-206
Wang, Zengfu II-77
Wei, Xin I-492
Weng, Li I-62
Wezel, Casper van II-157
Why, Yong Peng I-536
Wu, Bin II-359
Wu, Peng II-57
Wu, Pengcheng II-371
Wu, Qiang II-35
Wu, Shu-Min II-432
- Xu, Changsheng II-1, II-88
Xu, Chao II-479
- Yan, Chenggang I-51
Yan, Shuicheng II-1, II-465
Yan, Zhe I-413
Yang, Chunlei II-111
Yang, Jar-Ferr II-326
Yang, Jing II-274
Yeh, Wei-chang II-35
Yu, Chih-Chang II-187
Yu, Jen-Yu II-315
Yu, Like I-21
Yu, Meng-Chieh I-548
Yu, Nenghai I-403
- Zha, Zheng-Jun I-262, II-77
Zhang, Guoqing II-274
Zhang, Huanchen II-476
Zhang, Hui I-1
Zhang, Jinyu II-483
Zhang, Peng II-285
Zhang, Qing I-470
Zhang, Shanfeng II-359
Zhang, Tong II-65
Zhang, Xinxing II-88
Zhang, Yongdong I-21, I-51
Zhang, Zheng I-182
Zhao, Lili II-383
Zhao, Yi-Liang I-392
Zheng, Yan-Tao I-262, I-392
Zheng, Yu II-454
Zhou, Ning II-46
Zhou, Xiangdong I-392
Zhu, Guangyu II-1
Zhuang, Liansheng I-403