

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

Shipeng Li Abdulmotaleb El Saddik
Meng Wang Tao Mei
Nicu Sebe Shuicheng Yan
Richang Hong Cathal Gurrin (Eds.)

Advances in Multimedia Modeling

19th International Conference, MMM 2013
Huangshan, China, January 7-9, 2013
Proceedings, Part II



Springer

Volume Editors

Shipeng Li, Microsoft Research Asia, Beijing, China
E-mail: spli@microsoft.com

Abdulmoteleb El Saddik, University of Ottawa, ON, Canada
E-mail: elsaddik@uottawa.ca

Meng Wang, Hefei University of Technology, China
E-mail: eric.mengwang@gmail.com

Tao Mei, Microsoft Research Asia, Beijing, China
E-mail: tmei@microsoft.com

Nicu Sebe, University of Trento, Italy
E-mail: sebe@disi.unitn.it

Shuicheng Yan, National University of Singapore
E-mail: eleyans@nus.edu.sg

Richang Hong, Hefei University of Technology, China
E-mail: hongrc@hfut.edu.cn

Cathal Gurrin, Dublin City University, Ireland
E-mail: cgurrin@computing.dcu.ie

ISSN 0302-9743

ISBN 978-3-642-35727-5

DOI 10.1007/978-3-642-35728-2

Springer Heidelberg Dordrecht London New York

e-ISSN 1611-3349

e-ISBN 978-3-642-35728-2

Library of Congress Control Number: 2012954016

CR Subject Classification (1998): H.5.1, H.3.1, H.3.3-5, H.4.1, I.4, I.5, H.2.8, H.5.2, H.5.5

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

© Springer-Verlag Berlin Heidelberg 2013

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

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

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

Printed on acid-free paper

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

Preface

The 19th International Conference on Multimedia Modeling (MMM 2013) was held in Huangshan, China, during January 7–9, 2013, and hosted by the Hefei University of Technology (HFUT) at Hefei, China. MMM is a leading international conference for researchers and industry practitioners to share their new ideas, original research results, and practical development experiences from all multimedia-related areas.

It was a great honor for HFUT to host MMM 2013, one of the most longstanding multimedia conferences, in Huangshan, China. HFUT, located in the capital of Anhui province, is one of the key universities administrated by the Ministry of Education, China. Recently its multimedia-related research has been attracting increasing attention from the local and international multimedia community. The conference venue was the Huangshan International Hotel, located very close to Huangshan, which is well known for its scenery, sunsets, spectacularly shaped granite peaks, Huangshan pine trees, and unique views of the clouds from above. Furthermore, Huangshan is a UNESCO World Heritage Site and one of China's major tourist destinations. We hope that the choice of venue for MMM 2013 resulted in a memorable experience for all participants.

MMM 2013 featured a comprehensive program including three keynote talks, eight oral presentation sessions, one poster session, one demo session, seven special sessions, and the Video Browser Showdown. The 184 submissions from authors of 20 countries included a large number of high-quality papers in multimedia content analysis, multimedia signal processing and communications, and multimedia applications and services. We thank our 140-member Technical Program Committee who put in considerable effort in both reviewing papers and in providing valuable feedback to the authors. For the main conference, there were 111 submissions, each receiving at least three reviews. After the extensive reviewing process, the Program Chairs decided to accept 30 regular papers (27%) and 20 poster papers (18%). In total, 46 papers were accepted for seven special sessions after 53 were submitted, and 15 submissions were accepted for the demo session from a total of 22 submissions. Six teams competed in the Video Browser Showdown. The authors of accepted papers come from 16 countries. This volume of the conference proceedings contains the abstracts of three invited talks and all the regular, poster, special session, and demo papers, as well as special demo papers of the Video Browser Showdown (VBS). MMM 2013 included the following awards: the Best Paper Award, the Best Student Paper Award, two Best Demo Awards, and the VBS Competition Award, which were all sponsored by KAI Square.

The technical program is an important aspect but only has full impact if complemented by challenging keynotes. We were extremely pleased and grateful to have had three exceptional keynote speakers, Xian-Sheng Hua, Kiyoharu

Aizawa, and Ralf Steinmetz, accept our invitation and present interesting ideas and insights at MMM 2013.

We are also heavily indebted to many individuals for their significant contributions. We thank the MMM Steering Committee for their invaluable input and guidance on crucial decisions. We wish to acknowledge and express our deepest appreciation to the Honorary Chairs, Tat-Seng Chua, Phoebe Chen, and Wen Gao, the Local Organizing Chairs, Benoit Huet, Fei Wu, and Huaming Feng, the Special Session Chairs, Richang Hong and Changsheng Xu, the Demo Chairs, Ke Lu and Yi Yang, the VBS Chairs, Klaus Schöffmann and Werner Bailer, the Publicity Chairs, Kiyoharu Aizawa, Houqiang Li, Qingming Huang, Winston Hsu, and Xuelong Li, the Publication Chairs, Shuqiang Jiang and Cathal Gurrin, the Sponsorship Chairs, Yan-Tao Zheng, Shi-Yong Neo, Zhiwei Gu, and Qiong Liu, and last but not least the Webmaster, Xiaobin Yang. Without their efforts and enthusiasm, MMM 2013 would not have become a reality. Moreover, we want to thank our sponsors: Hefei University of Technology, National Natural Foundation of China, Beijing Ricoh Research Center, Microsoft Research Asia, FX Palo Alto Laboratory, and KAI Square Pte Ltd. We wish to thank all committee members, reviewers, session chairs, student volunteers, and supporters. Their contributions are much appreciated. Finally, we would also like to thank our local support team, Han Zhao, Xiaoping Liu, Yuanfa Zhu, Xiang Sun, Jianguo Jiang, Xuezhi Yang, Na Zhao, Yuetong Chen, Changzhi Luo, for their support and contribution to the conference organization.

January 2013

Tao Mei Nicu Sebe
Shuicheng Yan
Shipeng Li
Abdulmotaleb Ei Saddik
Meng Wang

Organization

MMM 2013 was organized by Hefei University of Technology, China.

MMM 2013 Organizing Committee

Honorary Co-chairs

Tat-Seng Chua	National University of Singapore, Singapore
Phoebe Chen	La Trobe University, Australia
Wen Gao	Peking University, China

General Co-chairs

Shipeng Li	Microsoft Research Asia, China
Abdulmotaleb Ei Saddik	University of Ottawa, Canada
Meng Wang	Hefei University of Technology, China

Program Co-chairs

Tao Mei	Microsoft Research Asia, China
Nicu Sebe	University of Trento, Italy
Shuicheng Yan	National University of Singapore, Singapore

Organizing Co-chairs

Benoit Huet	EURECOM, France
Fei Wu	Zhejiang University, China
Huaming Feng	Beijing Electronic Science and Technology Institute, China

Publicity Co-chairs

Kiyoharu Aizawa	University of Tokyo, Japan
Houqiang Li	University of Science and Technology of China, China
Qingming Huang	China Academy of Science, China
Winston Hsu	National Tai Wan University, Taiwan
Xuelong Li	XIOPM of Chinese Academy of Science, China

Sponsorship Co-chairs

Yantao Zheng	Google Corporation, USA
Shi-Yong Neo	Kai Square Co. Ltd., Singapore
Zhiwei Gu	Yahoo Corporation, USA
Qiong Liu	FX Palo Alto Laboratory, Inc., USA

Publication Co-chairs

Shuqiang Jiang	Institute of Computing, CAS, China
Cathal Gurrin	Dublin City University, Ireland

Special Session Co-chairs

Richang Hong	Hefei University of Technology, China
Changsheng Xu	Institute of Automation, China

Demo Co-chairs

Ke Lu	Chinese Academy of Science, China
Yi Yang	Carnegie Mellon University, USA

VBS Co-chairs

Klaus Schoeffmann	Klagenfurt University, Austria
Werner Bailer	Joanneum Research, Austria

Web Chair

Xiaobin Yang	Hefei University of Technology, China
--------------	---------------------------------------

US Liaisons

Qi Tian	University of Texas, San Antonio, USA
Alexander Hauptmann	Carnegie Mellon University, USA

Asian Liaisons

Chong-Wah Ngo	City University of Hong Kong, SAR China
Jialie Shen	Singapore Management University, Singapore

European Liaisons

Susanne Boll	University of Oldenburg, Germany
Alan Hanjalic	Delft University of Technology, The Netherlands

Technical Program Committee

Laurent Amsaleg	CNRS-IRISA, France
Xavier Anguera	Telefonica R&D, Spain
Yannis Avrithis	National Technical University of Athens, Greece
Bing-Kun Bao	CAS, China
Jenny Benois-Pineau	University of Bordeaux 1, France
Susanne Boll	University of Oldenburg, Germany
Laszlo Boszormenyi	Klagenfurt University, Austria
Liangliang Cao	IBM T.J. Watson Research, USA
Andrea Cavallaro	Queen Mary University of London, UK
Vincent Charvillat	University of Toulouse, France
Xiangyu Chen	National University of Singapore, Singapore
Gene Cheung	National Institute of Informatics, Japan
Liang-Tien Chia	Nanyang Technological University, Singapore
Wei-Ta Chu	National Chung Cheng University, Taiwan
Tat-Seng Chua	National University of Singapore, Singapore
Matthew Cooper	FX Palo Alto Laboratory, USA
Ajay Divakaran	Sarnoff Corporation, USA
Lingyu Duan	Peking University, China
Jianping Fan	University of North Carolina, USA
Yue Gao	National University of Singapore, Singapore
William Grosky	University of Michigan, USA
Cathal Gurrin	Dublin City University, Ireland
Martin Halvey	University of Glasgow, UK
Allan Hanbury	Technical University of Vienna, Austria
Andreas Henrich	University of Bamberg, Germany
Steven Hoi	Nanyang Technological University, Singapore
Richang Hong	Hefei University of Technology, China
Jun-Wei Hsieh	National Taiwan Ocean University, Taiwan
Winston Hsu	National Taiwan University, Taiwan
Benoit Huet	EURECOM, France
Wolfgang Hurst	Utrecht University, The Netherlands
Ichiro Ide	Nagoya University, Japan
Alejandro Jaimes	Yahoo!, USA
Rongrong Ji	Columbia University, USA
Yu-Gang Jiang	Columbia University, USA
Shuqiang Jiang	Chinese Academy of Sciences, China
Alexis Joly	INRIA, France
Mohan Kankanhalli	National University of Singapore, Singapore
Yoshihiko Kawai	NHK, Japan
Lyndon Kennedy	Yahoo! Research, USA
Yiannis Kompatsiaris	Informatics and Telematics Institute, Greece
Martha Larson	Delft University of Technology, The Netherlands
Duy-Dinh Le	National Institute of Informatics, Japan

Houqiang Li	University of Science and Technology of China, China
Chia-Wen Lin	National Tsing Hua University, Taiwan
Xiaobai Liu	University of California, Los Angeles, USA
Dong Liu	Columbia University, USA
Yan Liu	Hong Kong Polytechnic University, Hong Kong, SAR China
Zhu Liu	AT&T Laboratories, USA
Yuan Liu	Ricoh Software Research Center, China
Alexander Loui	Kodak Research Laboratories, USA
Guojun Lu	Monash University, Australia
Nadia Magnenat-Thalmann	University of Geneva, Switzerland
Jose Martinez	Universidad Autonoma de Madrid, Spain
Henning Mueller	HES-SO Valais, Switzerland
Francesco Natale	University of Trento, Italy
Chong Wah Ngo	City University of Hong Kong, Hong Kong, SAR China
Naoko Nitta	Osaka University, Japan
Noel O'Connor	Dublin City University, Ireland
Wei-Tsang Ooi	National University of Singapore, Singapore
Vincent Oria	New Jersey Institute of Technology, USA
Marco Paleari	EURECOM, France
Fernando Pereira	Instituto Superior Tecnico, Portugal
Guo-Jun Qi	University of Illinois at Urbana-Champaign, USA
Shin'ichi Satoh	National Institute of Informatics, Japan
Klaus Schoffmann	Klagenfurt University, Austria
Heng Tao Shen	University of Queensland, Australia
Jialie Shen	Singapore Management University, Singapore
Koichi Shinoda	Tokyo Institute of Technology, Japan
Mei-Ling Shyu	University of Miami, USA
Alan Smeaton	Dublin City University, Ireland
Cees Snoek	University of Amsterdam, The Netherlands
Yongqing Sun	NTT Cyber Space Laboratories, Japan
Jinhui Tang	Nanjing University of Science and Technology, China
Qi Tian	University of Texas at San Antonio, USA
Dian Tjondronegoro	Queensland University of Technology, Australia
Shingo Uchihashi	Carnegie Mellon University, USA
Xin-Jing Wang	Microsoft Research Asia, China
Zhiyong Wang	University of Sydney, Australia
Jingdong Wang	Microsoft Research Asia, China
Marcel Worring	University of Amsterdam, The Netherlands
Peng Wu	Hewlett-Packard, USA
Qiang Wu	University of Technology, Sydney, Australia
Xiao Wu	Southwest Jiaotong University, China

Feng Wu	Microsoft Research Asia, China
Changsheng Xu	Institute of Automation, Chinese Academy of Sciences, China
Keiji Yanai	University of Electro-Communications, Japan
Zheng-Jun Zha	National University of Singapore, Singapore
Zhongfei Zhang	State University of New York at Binghamton, USA
Yongdong Zhang	Institute of Computing Technology, CAS, China
Cha Zhang	Microsoft Research, USA
Roger Zimmermann	National University of Singapore, Singapore
Haojie Li	Dalian University of Technology, China

Additional Reviewers

Werner Bailer	Joanneum Research, Austria
Manfred del Fabro	Klagenfurt University, Austria
Frank Hopfgartner	DIA Laboratory, Technical University of Berlin, Germany
Mario Taschwer	Klagenfurt University, Austria
Wolfgang Weiss	Joanneum Research, Austria
Zhen Li	University of Illinois at Urbana-Champaign, USA
Ansgar Scherp	University of Koblenz-Landau, Germany
Makoto Okabe	University of Electro-Communications, Japan
Masaki Takahashi	NHK Science and Technology Research lab, Japan
Wen-Huang Cheng	Academia Sinica, Taiwan
Jiashi Feng	National University of Singapore, Singapore
Jitao Sang	Institute of Automation, Chinese Academy of Sciences, China
Congyan Lang	Beijing Jiaotong University, China
Si Liu	National University of Singapore, Singapore
Jian Cheng	Institute of Automation, Chinese Academy of Sciences, China
Jinqiao Wang	Institute of Automation, Chinese Academy of Sciences, China
Min-Hsuan Tsai	University of Illinois at Urbana, USA
Ming Yang	Northwestern University, USA
Peng Yang	Rutgers University, USA
Quan Fang	Institute of Automation, Chinese Academy of Sciences, China
Shiyang Lu	University of Sydney, Australia
Zhaowen Wang	University of Illinois at Urbana, USA
Weiqing Min	Institute of Automation, Chinese Academy of Sciences, China

Darui Li	University of Science and Technology of China, China
Yang Yang	The University of Queensland, Australia
Ming Yan	Institute of Automation, Chinese Academy of Sciences, China
Zhen Li	Dolby Laboratories, Inc., USA
Zhaoquan Yuan	Institute of Automation, Chinese Academy of Sciences, China
Yan Wang	Columbia University, USA
Xian-Ming Liu	University of Illinois at Urbana, USA
Pengfei Xu	Harbin Institute of Technology, China
Xiaoshuai Sun	Harbin Institute of Technology, China
Liujuan Cao	Harbin Engineering University, China

Special Session Co-chairs

Richang Hong	Hefei University of Technology, China
Changsheng Xu	Institute of Automation, China

Special Session Committee

Haojie Li	Dalian University of Technology, China
Shiguo Lian	Huawei Technologies, Co. Ltd., China
Yongqing Sun	NTT Cyber Space Laboratories, Japan
Jialie Shen	Singapore Management University, Singapore
Haiyan Miao	IHPC, A*STAR, Singapore
Liangliang Cao	IBM Watson Research Center, USA
Chang Wen Chen	SUNY at Buffalo, USA
Zhen Wen	IBM T.J. Watson Research Center, USA
Lu Fang	University of Science and Technology of China, China
Ngai-Man Cheung	Singapore University of Technology and Design, Singapore
Jingjing Fu	Microsoft Research Asia, China
Rongrong Ji	Columbia University, USA
Yue Gao	National University of Singapore, Singapore
Qingshan Liu	Nanjing University of Information Science and Technology, China
Wei-Ta Chu	National Chung Cheng University, Taiwan
Keiji Yanai	University of Electro-Communications, Japan
Bingkun Bao	Institute of Automation, Chinese Academy of Sciences, China
Jitao Sang	Institute of Automation, Chinese Academy of Sciences, China
Jinjun Wang	Epson Research and Development, USA

Best Paper Award Committee

Tat-Seng Chua	National University of Singapore, Singapore
Phoebe Chen	La Trobe University, Australia
Shipeng Li	Microsoft Research Asia, China
Abdulmotaleb Ei Saddik	University of Ottawa, Canada
Meng Wang	Hefei University of Technology, China
Tao Mei	Microsoft Research Asia, China
Nicu Sebe	University of Trento, Italy
Shuicheng Yan	National University of Singapore, Singapore

Sponsors List

Hefei University of Technology



National Natural Science Foundation of China



KAI Square Pte. Ltd



Microsoft Research Asia



Ricoh Beijing Software
Research Center



FX Palo Alto Laboratory



Google Inc.



Springer Publishing



Keynote 1: Perspective on Adaptive Video-Streaming

Prof. Dr.-Ing. Ralf Steinmetz

Department of Electrical Engineering and Information Technology
and Department of Computer Science in Technische Universität Darmstadt, Germany

Abstract. This talk covers perspectives on adaptive video streaming and how such techniques are essential for systems with heterogeneous devices. Adaptation is possible using flexible video coding techniques, such as the H.264 Scalable Video Coding (SVC). In this context, it is important to consider various aspects of the video coding system (interdependencies, quality layers, QoE, etc) as well of the delivery architectures (client server, P2P, connectivity, etc). The first part relates to quality adaptation algorithms that match the video quality with available local and system resources without any a-priori knowledge about those resources. Subsequently in the second part, mechanisms that use Quality of Experience (QoE) metrics to enhance its performance for the users will be shown. The decision of which SVC quality to choose is usually driven by QoS metrics, such as throughput. Instead, it will be presented how objective QoE of the different SVC qualities can be used in the decision process. The talk concludes by presenting the major further research activities in this research area.

Keynote 2: Multimedia FoodLog: Easiest Way to Capture and Archive What We Eat

Prof. Kiyoharu Aizawa

Department of Information and Communication Engineering
and Interfaculty Initiative of Information Studies of the University of Tokyo

Abstract. Eating is one of the most fundamental aspects of one's daily life, but at the same time, it is one of the most difficult aspects to manage by oneself. Recording what we eat is vital for our health care. We have been investigating the "FoodLog" multimedia food-recording tool, whereby users upload photos of their meals and a food diary is constructed by using image processing functions such as food image detection, dietary balance estimation, calory estimation etc. Foodlog is available in <http://www.foodlog.jp>, and to the best of our knowledge, it is currently the only publicly available multimedia food-recording application that makes use of image processing for dietary assessment. In addition to the PC-based interface, we have developed a few smartphone applications which makes easier to make detiled recording with the assist of image processing. In the talk, I would like to outline the current status of our FoodLog, and present various subjects on multimedia processing of foodlog.

Keynote 3: Towards Web-Scale Content-Aware Image Search

Dr. Xian-Sheng Hua

Microsoft Corporation

Abstract. In recent years, remarkable progress has been made towards large-scale content-aware image search. However, there is still a long way to go to bridge the two “gaps”: semantic gap and intention gap. Large-scale data brings us both challenges and opportunities to tackle these difficulties. In this presentation, we will review existing image search schemes and then focus on large-scale content-aware image search. We discuss the connections and differences among different large-scale CBIR techniques such as trees, clustering, hashing, graph and BoW, and then introduce a few exemplary scalable approaches including graph based search, color map based search, line sketch based search, and concept map based search. For each exemplary approach, we will discuss how to make it work for billions of images. The limitations will then be analyzed for these techniques, followed by introducing indexing and search schemes based on web-scale image content understanding. Connections between search and content understanding will be also discussed. And last we will talk about challenges and opportunities along this direction.

Table of Contents – Part II

Special Session Papers

Mobile-Based Multimedia Analysis

Quality Assessment on User Generated Image for Mobile Search Application	1
<i>Qiong Liu, You Yang, Xu Wang, and Liujuan Cao</i>	
2D/3D Model-Based Facial Video Coding/Decoding at Ultra-Low Bit-Rate	12
<i>Jun Yu, Zengfu Wang, and Yang Cao</i>	
Hierarchical Text Detection: From Word Level to Character Level	24
<i>Yanyun Qu, Weimin Liao, Shen Lu, and Shaojie Wu</i>	
Building a Large Scale Test Collection for Effective Benchmarking of Mobile Landmark Search	36
<i>Zhiyong Cheng, Jing Ren, Jialie Shen, and Haiyan Miao</i>	
Geographical Retagging	47
<i>Liujuan Cao, Yue Gao, Qiong Liu, and Rongrong Ji</i>	

Multimedia Retrieval and Management with Human Factors

Recompilation of Broadcast Videos Based on Real-World Scenarios	58
<i>Ichiro Ide</i>	
Evaluating Novice and Expert Users on Handheld Video Retrieval Systems	69
<i>David Scott, Frank Hopfgartner, Jinlin Guo, and Cathal Gurrin</i>	
Perfect Snapping	79
<i>Qingsong Zhu, Ling Shao, Qi Li, and Yaoqin Xie</i>	
Smart Video Browsing with Augmented Navigation Bars	88
<i>Manfred Del Fabro, Bernd Münzer, and Laszlo Böszörményi</i>	
Human Action Search Based on Dynamic Shape Volumes	99
<i>Hong-Ming Chen, Wen-Huang Cheng, Min-Chun Hu, Yan-Ching Lin, and Yung-Huan Hsieh</i>	

Video Retrieval Based on User-Specified Appearance and Application to Animation Synthesis	110
<i>Makoto Okabe, Yuta Kawate, Ken Anjyo, and Rikio Onai</i>	

Location Based Social Media

Landmark History Visualization	121
<i>Weiqing Min, Bing-Kun Bao, and Changsheng Xu</i>	
Discovering Latent Clusters from Geotagged Beach Images	133
<i>Yang Wang and Liangliang Cao</i>	

3D Video Depth and Texture Analysis and Compression

An Error Resilient Depth Map Coding Scheme Using Adaptive Wyner-Ziv Frame	143
<i>Xiangkai Liu, Qiang Peng, Xiao Wu, Lei Zhang, Xu Xia, and Lingyu Duan</i>	
A New Closed Loop Method of Super-Resolution for Multi-view Images	154
<i>Jing Zhang, Yang Cao, Zhigang Zheng, and Zengfu Wang</i>	
Fast Coding Unit Decision Algorithm for Compressing Depth Maps in HEVC	165
<i>Yung-Hsiang Chiu, Kuo-Liang Chung, Wei-Ning Yang, Yong-Huai Huang, and Chih-Ming Lin</i>	
Fast Mode Decision Based on Optimal Stopping Theory for Multiview Video Coding	176
<i>Hanli Wang, Yue Heng, Tiesong Zhao, and Bo Xiao</i>	
Inferring Depth from a Pair of Images Captured Using Different Aperture Settings	187
<i>Yujun Li, Oscar C. Au, Lingfeng Xu, Wenxiu Sun, and Wei Hu</i>	

Large-Scale Rich Media Search and Management in the Social Web

Multiple Instance Learning for Automatic Image Annotation	194
<i>Zhaoqiang Xia, Jinye Peng, Xiaoyi Feng, and Jianping Fan</i>	
Combining Topic Model and Relevance Filtering to Localize Relevant Frames in Web Videos	206
<i>Lei Yi, Haojie Li, and Shi-Yong Neo</i>	

A Lightweight Fingerprint Recognition Mechanism of User Identification in Real-Name Social Networks	217
<i>Haibin Cai, Zishan Qin, Yunyun Su, Junnan Tu, and Linhua Jiang</i>	
A Novel Binary Feature from Intensity Difference Quantization between Random Sample of Points	228
<i>Dongye Zhuang, Dongming Zhang, Jintao Li, and Qi Tian</i>	
Beyond Kmedoids: Sparse Model Based Medoids Algorithm for Representative Selection	239
<i>Yu Wang, Sheng Tang, Feidie Liang, YaLin Zhang, and Jintao Li</i>	
Improving Automatic Image Tagging Using Temporal Tag Co-occurrence	251
<i>Philip McParlane, Stewart Whiting, and Joemon Jose</i>	
Robust Detection and Localization of Human Action in Video	263
<i>Haojie Li, Fuming Sun, and Yue Guan</i>	

Multimedia Content Analysis Using Social Media Data

A Sparse Coding Based Transfer Learning Framework for Pedestrian Detection	272
<i>Feidie Liang, Sheng Tang, Yu Wang, Qi Han, and Jintao Li</i>	
Sampling of Web Images with Dictionary Coherence for Cross-Domain Concept Detection	283
<i>Yongqing Sun, Kyoko Sudo, Yukinobu Taniguchi, and Masashi Morimoto</i>	
Weakly Principal Component Hashing with Multiple Tables	293
<i>Haiyan Fu, Xiangwei Kong, Yanqing Guo, and Jiayin Lu</i>	
DUT-WEBV: A Benchmark Dataset for Performance Evaluation of Tag Localization for Web Video	305
<i>Haojie Li, Lei Yi, Yue Guan, and Hao Zhang</i>	
Clothing Extraction by Coarse Region Localization and Fine Foreground/Background Estimation	316
<i>Xiao Wu, Bo Zhao, Ling-Ling Liang, and Qiang Peng</i>	
Object Categorization Using Local Feature Context	327
<i>Tao Sun, Chunjie Zhang, Jing Liu, and Hanqing Lu</i>	
Statistical Multiplexing of MDFEC-Coded Heterogeneous Video Streaming	334
<i>Hang Zhang, Adarsh K. Ramasubramonian, Koushik Kar, and John W. Woods</i>	

Related HOG Features for Human Detection Using Cascaded Adaboost and SVM Classifiers	345
<i>Hong Liu, Tao Xu, Xiangdong Wang, and Yueqiang Qian</i>	
Face Recognition Using Multi-scale ICA Texture Pattern and Farthest Prototype Representation Classification	356
<i>Meng Wu, Jun Zhou, and Jun Sun</i>	
Detection of Biased Broadcast Sports Video Highlights by Attribute-Based Tweets Analysis	364
<i>Takashi Kobayashi, Tomokazu Takahashi, Daisuke Deguchi, Ichiro Ide, and Hiroshi Murase</i>	
Cross-Media Computing for Content Understanding and Summarization	
Temporal Video Segmentation to Scene Based on Conditional Random Fields	374
<i>Su Xu, Bailan Feng, and Bo Xu</i>	
Improving Preservation and Access Processes of Audiovisual Media by Content-Based Quality Assessment	385
<i>Peter Schallauer, Hannes Fassold, Albert Hofmann, Werner Bailer, and Stefanie Wechtitsch</i>	
Distribution-Aware Locality Sensitive Hashing	395
<i>Lei Zhang, Yongdong Zhang, Dongming Zhang, and Qi Tian</i>	
Cross Concept Local Fisher Discriminant Analysis for Image Classification	407
<i>Xinhang Song, Shuqiang Jiang, Shuhui Wang, Jinhui Tang, and Qingming Huang</i>	
A Weighted One Class Collaborative Filtering with Content Topic Features	417
<i>Ting Yuan, Jian Cheng, Xi Zhang, Qinshan Liu, and Hanqing Lu</i>	
Contextualizing Tag Ranking and Saliency Detection for Social Images	428
<i>Wen Wang, Congyan Lang, and Songhe Feng</i>	
Illumination Variation Dictionary Designing for Single-Sample Face Recognition via Sparse Representation	436
<i>Biao Wang, Weifeng Li, and Qingmin Liao</i>	

Efficient Extraction of Feature Signatures Using Multi-GPU Architecture	446
<i>Martin Kruliš, Jakub Lokoč, and Tomáš Skopal</i>	
Collaborative Tracking: Dynamically Fusing Short-Term Trackers and Long-Term Detector	457
<i>Guibo Zhu, Jinqiao Wang, Changsheng Li, and Hanqing Lu</i>	
A Real-Time Fluid Rendering Method with Adaptive Surface Smoothing and Realistic Splash	468
<i>Pengcheng Wang, Yong Zhang, Dehui Kong, and Baocai Yin</i>	
Multi-document Summarization Exploiting Semantic Analysis Based on Tag Cluster	479
<i>Jee-Uk Heu, Jin-Woo Jeong, Iqbal Qasim, Young-Do Joo, Joon-Myun Cho, and Dong-Ho Lee</i>	

Demo Session Papers

ShareDay: A Novel Lifelog Management System for Group Sharing	490
<i>Lijuan Marissa Zhou, Niamh Caprani, Cathal Gurrin, and Noel E. O'Connor</i>	
Helping the Helpers: How Video Retrieval Can Assist Special Interest Groups	493
<i>Frank Hopfgartner, Jinlin Guo, David Scott, Hongyi Wang, Yang Yang, Zhenxing Zhang, Lijuan Marissa Zhou, and Cathal Gurrin</i>	
Browsing Linked Video Archives of WWW Video	496
<i>Zhenxing Zhang, Cathal Gurrin, and Jinlin Guo</i>	
Multi-camera Egocentric Activity Detection for Personal Assistant	499
<i>Longfei Zhang, Yue Gao, Wei Tong, Gangyi Ding, and Alexander Hauptmann</i>	
Music Search Engine with Virtual Musical Instruments Playing Interface	502
<i>Mei Wang, Wei Mao, and Hai-Kiat Goh</i>	
Navilog: A Museum Guide and Location Logging System Based on Image Recognition	505
<i>Soichiro Kawamura, Tomoko Ohtani, and Kiyoharu Aizawa</i>	
Early Skip Mode Detection by Exploring Extra Skip Patterns for H.264 Coarse Grain Quality Scalable Video Coding	508
<i>Hao Zhang, Xiao Yu Zhu, and Xuan He</i>	

A Video Communication System Based on Spatial Rewriting and ROI Rewriting	511
<i>Hongtao Wang, Fangdong Chen, Bin Li, Dong Zhang, and Houqiang Li</i>	
<i>NEXT-Live: A Live Observatory on Social Media</i>	514
<i>Huanbo Luan, Dejun Hou, and Tat-Seng Chua</i>	
Online Boosting Tracking with Fragmented Model	517
<i>Dingcheng Shen, Hua Zhang, Yanbing Xue, Guangping Xu, and Zan Gao</i>	
Nonrigid Object Modelling and Visualization for Hepatic Surgery Planning in e-Health	521
<i>Suhuai Luo and Jiaming Li</i>	
Encoder/Decoder for Privacy Protection Video with Privacy Region Detection and Scrambling	525
<i>Feng Dai, Dongming Zhang, and Jintao Li</i>	
TVEar: A TV-tagging System Based on Audio Fingerprint	528
<i>Tao Jiang, Jiahong Li, Rihui Wu, and Kang Xiang</i>	
VTrans: A Distributed Video Transcoding Platform	532
<i>Zhe Ouyang, Feng Dai, Junbo Guo, and Yongdong Zhang</i>	
Fast ASA Modeling and Texturing Using Subgraph Isomorphism Detection Algorithm of Relational Model	535
<i>Feng Xue, Xiaotao Wang, Feng Liang, and Pingping Yang</i>	

Video Browser Showdown

An Approach for Browsing Video Collections in Media Production	538
<i>Werner Bailer, Wolfgang Weiss, Christian Schober, and Georg Thallinger</i>	
DCU at MMM 2013 Video Browser Showdown	541
<i>David Scott, Jinlin Guo, Cathal Gurrin, Frank Hopfgartner, Kevin McGuinness, Noel E. O'Connor, Alan F. Smeaton, Yang Yang, and Zhenxing Zhang</i>	
AAU Video Browser with Augmented Navigation Bars	544
<i>Manfred Del Fabro, Bernd Münzer, and Laszlo Böszörményi</i>	
NII-UIT-VBS: A Video Browsing Tool for Known Item Search	547
<i>Duy-Dinh Le, Vu Lam, Thanh Duc Ngo, Vinh Quang Tran, Vu Hoang Nguyen, Duc Anh Duong, and Shin'ichi Satoh</i>	

VideoCycle: User-Friendly Navigation by Similarity in Video Databases	550
<i>Christian Frisson, Stéphane Dupont, Alexis Moinet, Cécile Picard-Limpens, Thierry Ravet, Xavier Siebert, and Thierry Dutoit</i>	
Interactive Video Retrieval Using Combination of Semantic Index and Instance Search	554
<i>Hongliang Bai, Lezi Wang, Yuan Dong, and Kun Tao</i>	
Author Index	557

Table of Contents – Part I

Regular Papers

Multimedia Annotation I

Semi-supervised Concept Detection by Learning the Structure of Similarity Graphs	1
<i>Symeon Papadopoulos, Christos Sagonas, Ioannis Kompatsiaris, and Athena Vakali</i>	
Refining Image Annotation by Integrating PLSA with Random Walk Model	13
<i>Dongping Tian, Xiaofei Zhao, and Zhongzhi Shi</i>	
Social Media Annotation and Tagging Based on Folksonomy Link Prediction in a Tripartite Graph	24
<i>Majdi Rawashdeh, Heung-Nam Kim, and Abdulmotaleb El Saddik</i>	
Can You See It? Two Novel Eye-Tracking-Based Measures for Assigning Tags to Image Regions	36
<i>Tina Walber, Ansgar Scherp, and Steffen Staab</i>	

Multimedia Annotation II

Visual Analysis of Tag Co-occurrence on Nouns and Adjectives	47
<i>Yuya Kohara and Keiji Yanai</i>	
Verb-Object Concepts Image Classification via Hierarchical Nonnegative Graph Embedding	58
<i>Chao Sun, Bing-Kun Bao, and Changsheng Xu</i>	
Robust Semantic Video Indexing by Harvesting Web Images	70
<i>Yang Yang, Zheng-Jun Zha, Heng Tao Shen, and Tat-Seng Chua</i>	

Interactive and Mobile Multimedia

Interactive Evaluation of Video Browsing Tools	81
<i>Werner Bailer, Klaus Schoeffmann, David Ahlström, Wolfgang Weiss, and Manfred Del Fabro</i>	
Paint the City Colorfully: Location Visualization from Multiple Themes	92
<i>Quan Fang, Jitao Sang, Changsheng Xu, and Ke Lu</i>	

Interactive Video Advertising: A Multimodal Affective Approach	106
<i>Karthik Yadati, Harish Katti, and Mohan Kankanhalli</i>	
GPS Estimation from Users’ Photos	118
<i>Jing Li, Xueming Qian, Yuan Yan Tang, Linjun Yang, and Chaoteng Liu</i>	
Knowing Who You Are and Who You Know: Harnessing Social Networks to Identify People via Mobile Devices	130
<i>Mark Bloess, Heung-Nam Kim, Majdi Rawashdeh, and Abdulmotaleb El Saddik</i>	

Classification, Recognition and Tracking I

Hyperspectral Image Classification by Using Pixel Spatial Correlation	141
<i>Yue Gao and Tat-Seng Chua</i>	
Research on Face Recognition under Images Patches and Variable Lighting	152
<i>Wengang Feng</i>	
A New Network-Based Algorithm for Human Group Activity Recognition in Videos	163
<i>Gaojian Li, Weiyao Lin, Sheng Zhang, Jianxin Wu, Yuanzhe Chen, and Hui Wei</i>	
Exploit Spatial Relationships among Pixels for Saliency Region Detection Using Topic Model	174
<i>Guang Jiang, Xi Liu, JinPeng Yue, and Zhongzhi Shi</i>	

Classification, Recognition and Tracking II

Mining People’s Appearances to Improve Recognition in Photo Collections	185
<i>Markus Brenner and Ebroul Izquierdo</i>	
Person Re-identification by Local Feature Based on Super Pixel	196
<i>Cheng Liu and Zhicheng Zhao</i>	
An Effective Tracking System for Multiple Object Tracking in Occlusion Scenes	206
<i>Weizhi Nie, Anan Liu, Yuting Su, and Zan Gao</i>	

Ranking in Search

Image Search Reranking with Semi-supervised LPP and Ranking SVM	217
<i>Zhong Ji, Yanru Yu, Yuting Su, and Yanwei Pang</i>	
Co-ranking Images and Tags via Random Walks on a Heterogeneous Graph	228
<i>Lin Wu, Yang Wang, and John Shepherd</i>	
Social Visual Image Ranking for Web Image Search	239
<i>Shaowei Liu, Peng Cui, Huanbo Luan, Wenwu Zhu, Shiqiang Yang, and Qi Tian</i>	

Multimedia Representation

Fusion of Audio-Visual Features and Statistical Property for Commercial Segmentation	250
<i>Bo Zhang, Bailan Feng, and Bo Xu</i>	
Learning Affine Robust Binary Codes Based on Locality Preserving Hash	261
<i>Wei Zhang, Ke Gao, Dongming Zhang, and Jintao Li</i>	
A Novel Segmentation-Based Video Denoising Method with Noise Level Estimation	272
<i>Shijie Zhang, Jing Zhang, Zhe Yuan, Shuai Fang, and Yang Cao</i>	

Multimedia Systems

Blocking Artifact Reduction in DIBR Using an Overcomplete 3D Dictionary	283
<i>Cheolkon Jung, Licheng Jiao, and Hongtao Qi</i>	
Efficient HEVC to H.264/AVC Transcoding with Fast Intra Mode Decision	295
<i>Jun Zhang, Feng Dai, Yongdong Zhang, and Chenggang Yan</i>	
SSIM-Based End-to-End Distortion Model for Error Resilient Video Coding over Packet-Switched Networks	307
<i>Lei Zhang, Qiang Peng, and Xiao Wu</i>	
A Novel and Robust System for Time Recognition of the Digital Video Clock Using the Domain Knowledge	318
<i>Xinguo Yu, Tie Rong, Lin Li, and Hon Wai Leong</i>	
On Modeling 3-D Video Traffic	327
<i>M.E. Sousa-Vieira</i>	

Posters Papers

A Low-Complexity Quantization-Domain H.264/SVC to H.264/AVC Transcoder with Medium-Grain Quality Scalability	336
<i>Lei Sun, Zhenyu Liu, and Takeshi Ikenaga</i>	
Evaluation of Product Quantization for Image Search	347
<i>Wei-Ta Chu, Chun-Chang Huang, and Jen-Yu Yu</i>	
Rate-Quantization and Distortion-Quantization Models of Dead-Zone Plus Uniform Threshold Scalar Quantizers for Generalized Gaussian Random Variables	357
<i>Yizhou Duan, Jun Sun, and Zongming Guo</i>	
Flexible Presentation of Videos Based on Affective Content Analysis	368
<i>Sicheng Zhao, Hongxun Yao, Xiaoshuai Sun, Xiaolei Jiang, and Pengfei Xu</i>	
Dynamic Multi-video Summarization of Sensor-Rich Videos in Geo-Space	380
<i>Ying Zhang, He Ma, and Roger Zimmermann</i>	
Towards Automatic Music Performance Comparison with the Multiple Sequence Alignment Technique	391
<i>Chih-Chin Liu</i>	
Multi-frame Super Resolution Using Refined Exploration of Extensive Self-examples	403
<i>Wei Bai, Jiaying Liu, Mading Li, and Zongming Guo</i>	
Iterative Super-Resolution for Facial Image by Local and Global Regression	414
<i>Fei Zhou, Biao Wang, Wenming Yang, and Qingmin Liao</i>	
Stripe Model: An Efficient Method to Detect Multi-form Stripe Structures	425
<i>Yi Liu, Dongming Zhang, Junbo Guo, and Shouxun Lin</i>	
Saliency-Based Content-Aware Image Mosaics	436
<i>Dongyan Guo, Jinhui Tang, Jundi Ding, and Chunxia Zhao</i>	
Combining Visual and Textual Systems within the Context of User Feedback	445
<i>Leszek Kaliciak, Dawei Song, Nirmalie Wiratunga, and Jeff Pan</i>	
A Psychophysiological Approach to the Usability Evaluation of a Multi-view Video Browsing Tool	456
<i>Carmen Martinez-Peñaranda, Werner Bailer, Miguel Barreda-Ángeles, Wolfgang Weiss, and Alexandre Pereda-Baños</i>	

Film Comic Generation with Eye Tracking	467
<i>Tomoya Sawada, Masahiro Toyoura, and Xiaoyang Mao</i>	
Quality Assessment of User-Generated Video Using Camera Motion	479
<i>Jinlin Guo, Cathal Gurrin, Frank Hopfgartner, Zhenxing Zhang, and Songyang Lao</i>	
Multiscaled Cross-Correlation Dynamics on SenseCam Lifelogged Images	490
<i>N. Li, M. Crane, H.J. Ruskin, and Cathal Gurrin</i>	
Choreographing Amateur Performers Based on Motion Transfer between Videos	502
<i>Kenta Mizui, Makoto Okabe, and Rikio Onai</i>	
Large Scale Image Retrieval with Practical Spatial Weighting for Bag-of-Visual-Words	513
<i>Fangyuan Wang, Hai Wang, Heping Li, and Shuwu Zhang</i>	
Music Retrieval in Joint Emotion Space Using Audio Features and Emotional Tags	524
<i>James J. Deng and C.H.C. Leung</i>	
Analyzing Favorite Behavior in Flickr	535
<i>Marek Lipczak, Michele Trevisiol, and Alejandro Jaimes</i>	
Unequally Weighted Video Hashing for Copy Detection	546
<i>Jiande Sun, Jing Wang, Hui Yuan, Xiaocui Liu, and Ju Liu</i>	
Erratum to: Efficient HEVC to H.264/AVC Transcoding with Fast Intra Mode Decision	E1
<i>Jun Zhang, Feng Dai, Yongdong Zhang, and Chenggang Yan</i>	
Author Index	559

Quality Assessment on User Generated Image for Mobile Search Application

Qiong Liu¹, You Yang^{1,2}, Xu Wang¹, and Liujuan Cao⁴

¹ Department of Electronics & Information Engineering,
Huazhong University of Science & Technology, Wuhan, China

² Automation Department, Tsinghua University, Beijing, China

³ Department of Computer Science, City University of Hong Kong, Hong Kong

⁴ College of Computer Science and Technology,
Harbin Engineering University, Harbin, China

{dr.qiongliu,youyang.sayu,wangxu.cise,caolijuan}@gmail.com

Abstract. Quality specified image retrieval is helpful to improve the user experiences in mobile searching and social media sharing. However, the model for evaluating the quality of the user generated images, which are popular in social media sharing, remains unexploited. In this paper, we propose a scheme for quality assessment on user generated image. The scheme is formed by four attribute dimensions, including intrinsic quality, favorability, relevancy and accessibility of images. Each of the dimensions is defined and modeled to pool a final quality score of a user generated image. The proposed scheme can reveal the quality of user generated image in comprehensive manner. Experimental results show that the scores obtained by our scheme have high correlation coefficients with the benchmark data. Therefore, our scheme is suitable for quality specified image retrieval on mobile applications.

Keywords: image quality assessment, user generated content, image retrieval, mobile search.

1 Introduction

User generated contents (UGC) in social media, especially images or photos [1, 2], provide a new way for web publishing and media production circle [3]. This kind of images is re-editable, accessible and affordable for general public in communication and mind exchange on social networks. Due to its conveniences, billions of user generated images (UGI) are published on web. For example, Facebook has 100 billion of images uploads from end-users, while Flickr has 60 billion and with 20% annual increase [4, 5]. The exploding volume of UGI brings a great difficulty to traditional image retrieval system [6–17], especially to the bandwidth and capacity constrained mobile or portable search platform. How can we improve the quality of UGI becomes a new paradigm for current and future mobile search and other applications [18–20].

The research on the quality assessment of mobile search can be generally divided into post- and prior-assessment. For the case of post-assessment, it mainly

focuses on the re-ranking of the retrieval results [21–23]. The post-assessment is performed on the retrieval results, which is difficult to be applied to the mobile search applications. On the other hand, prior-assessment can provide quality specification as user input to mobile search applications. The retrieval results can be convergent to a small range with quality constrained, and this UGI retrieval is benefit to low bandwidth communication and user experiences. However, the re-search on prior-assessment of UGI retrieval to mobile search remains unexploited so far.

In this paper, we propose a scheme on UGI quality assessment for mobile search applications. With this scheme, quality specification can be used as a prior constrain and input to a retrieval system. The scheme is benefit for mobile search applications due to great bandwidth savings, and the retrieval results are benefit for user experiences. The rest of this paper is organized as following. Section 2 provides descriptions on the related works. After that, the dimensions of UGI quality assessment are defined and analyzed in details in Section 3. Section 4 is for experimental results and discussions. Finally, our work is concluded in Section 5.

2 Related Work

Actually, this field is related to social presentation in social media, image quality assessment in signal processing, data quality assessment in work flow optimization and preference-based filtering in recommender system.

2.1 Social Presentation

The key reason why people decide to create an image and publish it in general public is the wish to present themselves in cyberspace [24]. Usually, such a presentation is done through self-presence and self-disclosure [25]. Therefore, the contents (i.e., UGI) the end-users posted on web are able to reveal their opinions. To this end, the quality of web published image is important to the UGI provider for self-presence to his friend-circle and the general public.

2.2 Image Quality Assessment

Image quality is a characteristic of a processed image that measures the perceived image degradation which typically comparing to its original version. This characteristic is important in imaging system and signal processing to quantify the distortions or artifacts [26]. Distortion measurement and score pooling is the basic approach for traditional image quality assessment models, and it has been studied in three different manners, including full-reference, reduced-reference and no-reference, according to the existence and completeness of the original image [26, 27]. However, traditional image quality assessment approaches for imaging system or signal processing are not suitable for UGI from social media.

2.3 Data Quality Assessment

Information or data consist of propositions that reflect reality, and data/information quality is the fitness for use by data consumers [28, 29]. Data are necessary and useful for daily work, but the exploding data volume can confuse data consumer and dull the work efficiency. In this case, data quality is in need to optimize the work flow and improve the work efficiency. The background of the research on data quality assessment is very close to the UGI quality assessment. The core structure of data quality assessment is the dimension design for fitness measurement [28]. This dimension structure will be helpful for the research of UGI quality assessment.

2.4 Recommender System

Recommender system seeks to predict the *rating* or *preference* that a user would give to an item (such as music, books, or movies) or social element (e.g. people or groups) they had not yet considered, using a model built from the characteristics of an item (content-based approaches) or the user's social environment [30, 31]. Actually, recommender system is a retrieval system with user specifications. The specifications, such as the favorite type of goods, selection habit, location and other useful information, can be obtained by historical purchase data of this registered user. Methods for modeling and analyzing the historical data are summarized as preference-based filtering methods, which include content-based recommendation, collaborative recommendation, and hybrid approaches [31].

3 The Proposed Scheme

In this paper, we evaluate the quality of UGI by a multi-dimensional approach. In this section, we first define the dimensions of UGI quality assessment. After that, each of the dimensions will be discussed in details.

3.1 Dimensions of User Generated Image Quality Assessment

Based on the discussions in Section 2, it can be found that the following attributes are indispensable when evaluating the quality of UGI.

(1) Intrinsic quality. Intrinsic quality of UGI denotes that image has the quality in its own right, for example the perceived quality, or aesthetic quality.

(2) Favorability. The UGI published on social networks can be reviewed by end-users, and these end-users can be friends or strangers to the image provider. If this image is favorable, it will receive positive comments, and even be recommended to other end-users or kept as favorites by the reviewer.

(3) Relevancy. Relevancy is a parameter that describe the degree how the current image matches the reviewer's requirement. The resultant images that with the characteristics of the input keywords or images are sorted by relevancy and feed backed to reviewer.

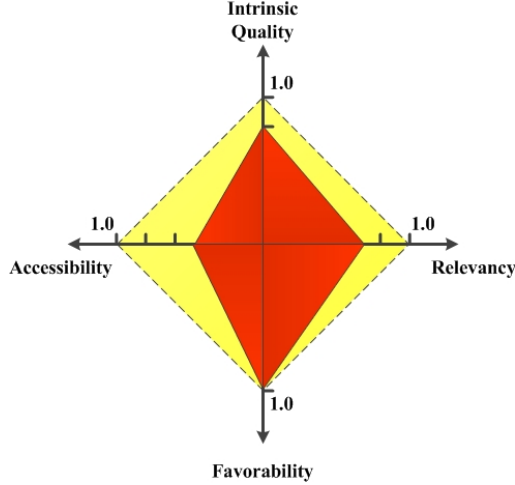


Fig. 1. Dimensions of user generated image quality assessment. The score for each dimension is normed between [0,1], and the scores will be pooled and normed to obtain the final score of the user generated image quality.

(4) Accessibility. Only when the retrievable image can be fetched easily and smoothly, the quality of this image is meaningful to the end-user.

As shown by Figure 1, the aforementioned attributes are important for UGI quality assessment, and are proposed to be the four dimensions to form the framework of quality assessment. The final score S of the UGI quality can be pooled as following

$$S = \frac{1}{2} \left(\frac{(S_I + S_F)S_A}{2} + \frac{(S_I + S_F)S_R}{2} \right) \\ = \frac{(S_I + S_F)(S_A + S_R)}{4}$$

where S_I , S_F , S_A and S_R stands for the score of intrinsic quality, favorability, accessibility and relevancy respectively.

3.2 Intrinsic Quality Assessment

As mentioned above, the intrinsic quality of UGI is the characteristic that measures the perceived image degradation. In most cases, a counterpart image is needed in the procedure of quality evaluation, and the image is usually the original one assumed with perfect quality. However, there is not counterpart image for UGI in quality evaluation. UGI is usually captured by end-user, simply edited and then published in general public. The degradation of UGI may come from two possible operations, which one is image compression (e.g. quantization losses by JPEG format encoding), and the other one is manual re-editing by users. As for the degradation caused by manual re-editing, it is the art re-generation on

the original image that aesthetics and emotions may be involved in the procedure of re-editing. In this case, it is very difficult to effectively evaluate this kind of art images [32]. On the other hand, the degradation caused by compression is a kind of impairment on signals. No-reference image quality assessment method is an effective approach to predict the perceived quality of this kind of images.

We adopt the no-reference image quality assessment model in [33] in our scheme, because this model is suitable for JPEG compressed images and is verified by LIVE image database [34]. This model can be described as

$$S_I = \alpha + \beta B^{\gamma_1} A^{\gamma_2} Z^{\gamma_3}$$

where α , β , γ_1 , γ_2 and γ_3 are model parameters, and

$$B = \frac{B_h + B_v}{2}, A = \frac{A_h + A_v}{2}, Z = \frac{Z_h + Z_v}{2}$$

where

$$B_* = \frac{1}{M(\lfloor N/8 \rfloor - 1)} \sum_{i=1}^M \sum_{j=1}^{\lfloor N/8 \rfloor - 1} |d_*(i, 8j)|$$

$$A_* = \frac{1}{7} \left[\frac{8}{M(N-1)} \sum_{i=1}^M \sum_{j=1}^{N-1} |d_*(i, j)| - B_* \right]$$

$$Z_* = \frac{1}{M(N-2)} \sum_{i=1}^M \sum_{j=1}^{N-2} z_*(m, n)$$

where $*$ can be one of h and v , indicating horizontal and vertical pixels.

3.3 Favorability Measurement

Comments to the UGI are able to describe the favorability of this image. The comment usually shows the opinion or attitude of the reviewer to the image, indicating sentiment such as love, like, happy, noncommittal, dislike, disgust, and so on. These sentiments can be used to describe the favorability of the images.

The problem of comments based favorability measurement is how to convert the natural words based comments into a numerical parameter describing the degree of favorite for an image. There is a huge gap between natural words and the numerical parameters. Fortunately, these comments written by natural words can be processed by sentiment analysis or opinion mining methods. Sentiment analysis or opinion mining refers to the application of natural language processing, computational linguistics, and text analytics to identify and extract subjective information in source materials [35]. With the help of sentiment analysis, the favorability of the UGI is measurable.

The problem for comment-based favorability evaluation is the number of types for mood. Traditionally, there are only three types of moods, including positive, noncommittal and negative (in some cases, there are two types without noncommittal) [35]. In order to evaluate the favorability by user comments, we first propose a 5-score method by means of mood. Comments with sentiments are classified into different levels, corresponding to different score value, as given by Table 1. The key words of mood level are the typical words that represent a group of mood words. Some examples are given in Table 2. According to this table, a score can be easily obtained by one comment if one mood word appearing in this comment.

Table 1. The look-up table for 5-score and mood

Score	Description	Mood	Mood Keywords
5	best	positive	surprise, best, extra-, ultra-
4	better	positive	much, very
3	good	positive	happy, good, alive, love, interested, positive, strong
2	fair	noncommittal	open
1	bad	negative	anger, disgust, fear, sadness

Table 2. Selected mood words for different sentiments

Mood Keywords	Words
open	understanding, confident, reliable, easy, free, sympathetic
happy	great, gay, joyous, lucky, fortunate, delighted, overjoyed, gleeful
alive	playful, courageous, energetic, liberated, optimistic, provocative
good	calm, peaceful, at ease, comfortable, pleased, encouraged, clever
love	loving, considerate, affectionate, sensitive, tender, devoted
interested	concerned, affected, fascinated, intrigued, absorbed, inquisitive
positive	eager, keen, intent, anxious, inspired, determined, excited
angry	irritated, enraged, hostile, insulting, sore, annoyed, upset, hateful
depressed	lousy, disappointed, discouraged, ashamed, powerless, diminished
sad	tearful, sorrowful, pained, grief, anguish, desolate, desperate

Usually, there are more than one comment (suppose the number is $n \geq 1$) posted by UGI reviewers. In this case, the favorability can be evaluated by the comment based score in terms of average score as following

$$S_F = \begin{cases} 0 & n = 0 \\ \frac{1}{5n} \sum_{i=1}^n c_i & n \geq 1 \end{cases}$$

where c_i is the comment based score for the i th comment.

3.4 Relevancy Measurement

The relevancy measurement method evaluates the relevancy between the image search results and the ones really wanted by users. According to the image retrieval algorithm, the images at lower ranking order have a less relevancy with the search keywords. Therefore, the relevancy can be measured by the ranking order of the obtained results. Assume that a particular image is ranked as k order in n results, and then the relevancy of this image can be calculated as

$$S_R = 1 - \frac{k}{n}$$

3.5 Accessibility Measurement

Accessibility represents whether this image can be obtained by users. For some UGI in general public, the owner preserves the copyright and sometimes limits the downloading and re-editing on these images. Therefore, the score of accessibility measurement is a binary parameter, which can be described as

$$S_A = \begin{cases} 1 & \text{obtainable} \\ 0 & \text{otherwise} \end{cases}$$

4 Experiments and Discussions

4.1 Experiment Arrangements

We download UGIs and the corresponding comments from Flickr by searching the keywords cat, street, dog, sea and car. The test images are randomly selected from the searching results, and the total number is 25. Figure 2 provides some of the test images. Both subjective and objective experiments are involved in our experiments. In subjective experiments, image reviewers are invited to offer their subjective score to each of the test images. The results of subjective experiments are used as benchmark data for later results of objective experiments. In objective experiments, the proposed scheme is performed and compared to the benchmark results.

4.2 Subjective Experiments

The subjective experiments are arranged strictly according to the procedure described in [36], including the participants selection, environment settings and display settings. There are 10 participants (2 female and 8 male) involved in the subjective experiments, ageing from 21 to 32 with different academic level and professional background, and all of them were not involved in similar subjective experiments in the past three months. The experimental data processing techniques described in [34, 36] are adopted for us to obtain benchmark data.



Fig. 2. Examples of test UGIs fetched from Flickr website

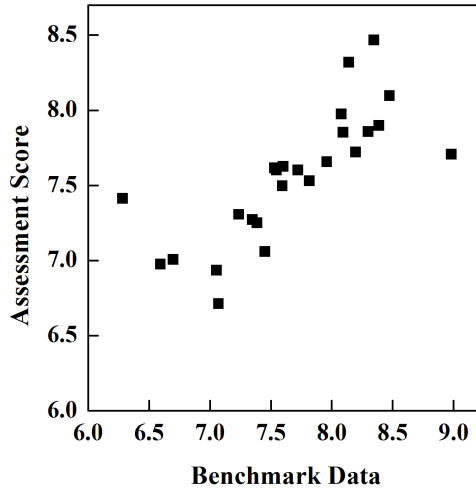


Fig. 3. Linear corresponding relationships between assessment scores and benchmark data

Table 3. The look-up table for 5-score and mood

Correlation type	Correlation parameter	Significance parameter
Pearson	0.76161	9.76343×10^{-6}
Spearman	0.88902	2.88471×10^{-6}
Kendall	0.68448	1.67552×10^{-6}

4.3 Objective Experiments

Besides the subjective experiments, we also use our proposed scheme to the involved 25 UGIs to obtain the objective experimental results. These results are compared to the benchmark data obtained above to verify the accuracy of quality prediction. The correlation between the predicted data and benchmark data can be used to reveal the performance of our proposed scheme, and the comparison results are summarized in Figure 3 and Table 3.

Our scheme evaluate the quality of UGI in four dimensions, including intrinsic quality, favorability, relevancy and accessibility, which reveal the user experiences in image search, especially the mobile search. Therefore, the scheme can predict the quality of UGI efficiently. As given by Figure 3, the score predicted by our scheme has linear relationship to the benchmark data which is benefit for image retrieval system in describing the user specifications. In order to show the performance well, the correlation parameter in types of Pearson, Spearman and Kendall are given by Table 3. Higher value of correlation parameter and lower value of significance parameter indicates better performance on score prediction. The results in Table 3 show that our scheme can predict the UGI quality efficiently.

5 Conclusions

In this paper, we propose a scheme for quality assessment of user generated image. The scheme is formed by four dimensions, including intrinsic quality, favorability, relevancy and accessibility of the user generated image. These dimensions are the most important attributes in evaluating the quality of image that published in general public. Experiments are arranged with subjective and objective ones, and the results show that the objective scores predicted by our scheme have high correlation parameter to the subjective scores.

Acknowledgments. This work was supported by NSFC (Grant No. 61170194 and 61202301).

References

1. Ji, R., Duan, L., Chen, J., Yao, H., Yuan, J., Rui, Y., Gao, W.: Location discriminative vocabulary coding for mobile landmark search. *International Journal of Computer Vision* 96(3), 290–314 (2012)

2. Ji, R., Gao, Y., Zhong, B., Yao, H., Tian, Q.: Mining flickr landmarks by modeling reconstruction sparsity. *ACM Transactions on Multimedia Computing, Communications, and Applications* 7(1), 1–22 (2011)
3. Daugherty, T., Eastin, M.S., Bright, L.: Exploring consumer motivations for creating user-generated content. *Journal of Interactive Advertising* 8(2), 16–25 (2008)
4. Parfeni, L.: Flickr boasts 6 billion photo uploads. Technical report
5. Cha, M., Kwak, H., Rodriguez, P., Ahn, Y.Y., Moon, S.: I tube, you tube, everybody tubes: Analyzing the world’s largest user generated content video system. In: *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement, IMC 2007*, pp. 1–14 (2007)
6. Gao, Y., Tang, J., Hong, R., Yan, S., Dai, Q., Zhang, N., Chua, T.: Camera constraint-free view-based 3D object retrieval. *IEEE Transactions on Image Processing* 21(4), 2269–2281 (2012)
7. Gao, Y., Wang, M., Zha, Z., Tian, Q., Dai, Q., Zhang, N.: Less is more: Efficient 3D object retrieval with query view selection. *IEEE Transactions on Multimedia* 11(5), 1007–1018 (2011)
8. Gao, Y., Wang, M., Tao, D., Ji, R., Dai, Q.: 3D object retrieval and recognition with hypergraph analysis. *IEEE Transactions on Image Processing* 21(9), 4290–4303 (2012)
9. Wang, M., Yang, K., Hua, X.S., Zhang, H.J.: Towards a relevant and diverse search of social images. *IEEE Transactions on Multimedia* 12(8), 829–842 (2010)
10. Wang, M., Hua, X.S.: Active learning in multimedia annotation and retrieval: A survey. *ACM Trans. Intell.* 2(2), 10:1–10:21 (2011)
11. Wang, M., Ni, B., Hua, X.S., Chua, T.S.: Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Comput. Surv.* 44(4), 25:1–25:24 (2012)
12. Shen, J., Pang, H., Wang, M., Yan, S.: Modeling concept dynamics for large scale music search. In: *SIGIR*, pp. 455–464 (2012)
13. Shen, J., Tao, D., Li, X.: Modality mixture projections for semantic video event detection. *IEEE Trans. Circuits Syst. Video Techn.* 18(11), 1587–1596 (2008)
14. Shen, J., Shepherd, J., Cui, B., Tan, K.L.: A novel framework for efficient automated singer identification in large music databases. *ACM Trans. Inf. Syst.* 27(3) (2009)
15. Zha, Z., Wang, M., Zheng, Y.T., Yang, Y., Hong, R., Chua, T.S.: Interactive video indexing with statistical active learning. *IEEE Transaction on Multimedia* 14(1), 17–27 (2012)
16. Zha, Z.J., Yang, L., Mei, T., Wang, M., Wang, Z., Chua, T.S., Hua, X.S.: Visual query suggestion: Towards capturing user intent in internet image search. *ACM Transactions on Multimedia Computing, Communications and Applications* 6(3) (2010)
17. Zha, Z.J., Yang, L., Mei, T., Wang, M., Wang, Z.: Visual query suggestion. In: *ACM Conference on Multimedia*, pp. 15–24 (2009)
18. Liu, Q., Yang, Y., Ji, R., Gao, Y., Yu, L.: Cross-view down/up-sampling method for multiview depth video coding. *IEEE Signal Processing Letters* 19(5), 295–298 (2012)
19. Yang, Y., Dai, Q.: Contourlet-based image quality assessment for synthesised virtual image. *Electronics Letters* 46(7), 492–494 (2010)
20. Wang, X., Yu, M., Yang, Y., Jiang, G.: Research on subjective stereoscopic image quality assessment. In: *Proceedings of SPIE: Multimedia Content Access: Algorithms and Systems III* (2009)

21. Tian, X., Yang, L., Wang, J., Wu, X., Hua, X.S.: Bayesian visual reranking. *IEEE Transactions on Multimedia* 13(4), 639–652 (2011)
22. Tian, X., Tao, D.: Visual reranking: From objectives to strategies. *IEEE MultiMedia* 18(3), 12–21 (2011)
23. Gao, Y., Wang, M., Zha, Z., Shen, J., Li, X., Wu, X.: Visual-textual joint relevance learning for tag-based social image search. *IEEE Transactions on Image Processing* (in press), doi:10.1109/TIP.2012.2202676
24. Schau, H.J., Gilly, M.C.: We are what we post? self presentation in personal web space. *Journal of Consumer Research* 30(3), 385–404 (2003)
25. Erving, G.: *The Presentation of Self in Everyday Life*, pp. 17–25. The Overlook Press, New York (1959)
26. Wang, Z., Sheikh, H.R., Bovik, A.C.: *Objective Video Quality Assessment*, pp. 1041–1078. CRC Press (2003)
27. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13(4), 600–612 (2004)
28. Wang, R.Y., Strong, D.M.: Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems* 12(4), 5–33 (1996)
29. Leea, Y.W., Strongb, D.M., Kahnc, B.K., Wang, R.Y.: Aimq: a methodology for information quality assessment. *Information & Management* 40(2), 133–146 (2002)
30. Ricci, F., Rokach, L., Shapira, B.: Introduction to recommender systems handbook. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) *Recommender Systems Handbook*, pp. 1–35. Springer US (2011)
31. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering* 17(6), 734–749 (2005)
32. Joshi, D., Datta, R., Fedorovskaya, E., Luong, Q.T., Wang, J., Li, J., Luo, J.: Aesthetics and emotions in images. *IEEE Signal Processing Magazine* 28(5), 94–115 (2011)
33. Wang, Z., Sheikh, H., Bovik, A.: No-reference perceptual quality assessment of jpeg compressed images. In: 2002 International Conference on Image Processing, vol. 1, pp. I-477 – I-480 (2002)
34. Sheikh, H., Sabir, M., Bovik, A.: A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing* 15(11), 3440–3451 (2006)
35. Liu, B.: Opinion mining and sentiment analysis. In: Carey, M.J., Ceri, S. (eds.) *Web Data Mining. Data-Centric Systems and Applications*, pp. 459–526. Springer, Heidelberg (2011)
36. Methodology for the subjective assessment of the quality of television pictures. ITU-R BT.500-11 (2002)

2D/3D Model-Based Facial Video Coding/Decoding at Ultra-Low Bit-Rate

Jun Yu, Zengfu Wang, and Yang Cao

University of Science and Technology of China, Hefei, Anhui, P.R. China
{harryjun,forrest,zfwang}@ustc.edu.cn

Abstract. For facial video coding/decoding in mobile communication, a 2D/3D model-based system was proposed: (1) Online appearance model and cylinder head model were combined to track 3D facial motion in particle filter; (2) 3D facial animation was produced combining parameterized model and muscular model; (3) 3D hair was synthesized by hair detection and 3D hair model; (4) 3D coding/decoding of foreground and 2D coding/decoding of background were stitched. At ultra-low bit-rate, the object experiment confirmed the advantage between coding efficiency and decoding quality of it, and the between subjects experiment indicated the suitability of subjective face identification by it.

1 Introduction

Currently, video communication playing a more and more important role in mobile electronics products. Especially, the coding scheme is a critical factor as vast data need to be transmitted through channel with limited capacity.

Traditional video coding treats image as pixel-by-pixel or transformed coefficients, and uses statistical characteristics to remove redundancy. However, 3D model-based video coding(MBC)[1] treats image as structural signal. The workflow is: in encoder, the object in input image is identified, and a 3D model is adapted to it. Then parameters describing the object are coded and transmitted. In decoder, the decoder's model is modified and synthesized by decoded parameters. In following frames, parameters describing the change of model are transmitted. MBC has high efficiency while has not square artifact, and is mainly applied to facial video(MBC-Face) for convenience. The relevant technologies have also been investigated in video driven facial animation[2], and mesh coding in MPEG-4, which is mostly a choice of compression parameters and the tracking method is not standardized. Currently, MBC-Face has following progresses.

For facial motion tracking, feature-based approaches[3] usually suffer from drifting. To overcome it, some measures should be utilized[4]. Appearance-based approaches are generally more robust than the above, and can be performed deterministically or statistically. The former[5] uses object-specified model, thus is sensitive to lighting and expression. The latter can be applied offline or online. The offline technique[6,7] cannot adapt to new condition which is not reflected in training database. The online technique(OAM)[8,9] learns appearance progressively, thus is more flexible than offline technique. To improve the flexibility

of OAM and reduce the influence of lighting and person dependence, one way is to exploit OAM[10,11], another way is to add extra information[12]. Especially, the global head motion can be got by simple geometric head models, e.g., cylinder[13], by minimizing the difference between observation and model, and can be described by only 6 parameters, thus results in robust tracking. In comparison[14] of different geometric models, the simplest ellipsoid model showed better performance than the generic shape model. This result comes from the fact that detailed information about face in latter imposes wrong prior that is apt to lose tracking. They do not require learning so as to get person independence, and they are robust to large pose change as the use of whole area of head in the image. Furthermore, for motion filtering in tracking, particle filter[11] was proved to be robust to 2D facial motion tracking, but has 3 problems, i.e., search blindness, computational effort and sample degeneration, for 3D tracking.

For facial animation, parameterized model[15] control shape and motion of facial model by parameterized equations. It is fast and effective, but hard to design parameters for high reality; Learning model[16] renders desired animation by the variations of shape/texture from 3D face dataset, but the cost is expensive; Physical model[17] simulates skin, tissue, and muscle by multi-layer deformed meshes, but it is computation-intensive to get high reality; Muscular model[18] simulates muscle motion from anatomy. It is suitable for facial animation intuitively, but the reality and computational effort need to be improved.

Hair affects the reality of facial animation largely[19], but its detection, analysis, and synthesis have not been studied in the MBC-Face community.

Video contain foreground(face and hair) and background[20]. However, background cannot be described by 3D model as the complexity, thus background and foreground need to be stitched for successful application.

In this paper, a 2D/3D mixed facial coding/decoding system is proposed. Compared to current MBC-Face system[20] and 2D scheme, e.g., H.264, our system: 1)includes 3D face, hair and 2D background; 2)has better rate/distortion performance at ultra-low bit-rate; 3)has the suitability of subjective face identification at ultra-low bit-rate.

2 3D Facial Model and Facial Model Adaptation

In encoder, we use CANDIDE3[6] model (fig. 1(a)). The shape of it is described by the vector g which is the concatenation of the 3D coordinates of all vertices:

$$g = \bar{g} + S\sigma + A\alpha \quad (1)$$

\bar{g} is standard shape. S , A are shape and animation units. σ , α are shape and animation parameters. Moreover, global motion parameters[9] are defined as:

$$h = [\theta_x, \theta_y, \theta_z, t_x, t_y, t_z]^T \quad (2)$$

Therefore, facial motion parameters are defined as:

$$b = [h^T, \sigma^T, \alpha^T]^T = [\theta_x, \theta_y, \theta_z, t_x, t_y, t_z, \sigma^T, \alpha^T]^T \quad (3)$$

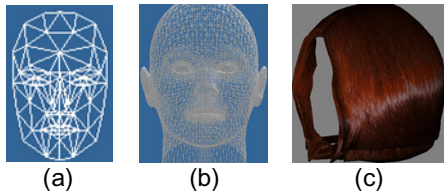


Fig. 1. (a) CANDIDE3 in encoder. (b) Alice in decoder. (c) Hair model in decoder.

In decoder, we use Alice[21] model(fig. 1(b)) including skin, eyes, teeth, etc. Especially, hair model(fig. 1(c)) is added to increase reality.

Facial model adaptation is applied to reconstruct an individual model by deforming a generic model with the first frame containing frontal face as follows.

In encoder: 1)facial feature points are got by AAM[7]; 2)global motion parameters are got by several feature points[22]; 3)displacements of dominance vertices are got by least square fitting of all feature points; 4)displacements of other vertices are got by radial based interpolation(RBF).

In decoder: 1)displacements of dominance vertices are got by facial motion parameter from encoder; 2)displacements of other vertices are got by RBF.

3 Encoder

3.1 OAM-Based Facial Motion Tracking

Measurements are extracted from geometrically normalized facial image(GNFI, fig. 2)[6]. Refer to [6] for detail that how to get GNFI.

The measurements include two parts, the first is pixel color value. As illumination ratio image[12] is independent to surface albedo and high frequencies of it are less influenced by lighting, and Gabor wavelet can capture local structure information corresponding to scale, direction, etc. The process of extracting the second measurement is: we get illumination ratio image between GNFI of first frame and current frame, then extract Gabor wavelet coefficients from it, and only the magnitudes of coefficients are used for resisting noise.

Assuming y_t is the concatenation of first measurement at time t , it is modeled as a GMM variable with 3 components[8]. The OAM A_t represents the stochastic

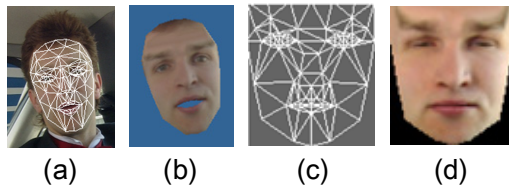


Fig. 2. The process of getting geometrically normalized facial image

process of all observations until time $t-1$: $y_{1:t-1}$. It is updated when b_t, y_t are got as [11] shows. Assuming G_t is the concatenation of second measurement at time t , it is also modeled as a GMM variable. The definition and corresponding OAM are similar to above. Then, the two measurements are fused by $p(y_t/b_t) \cdot p(G_t/b_t)$, which is set as particle weight in following section.

Motion Filtering by Particle Filter. Local optimization is added to particle filter for reducing the search blindness and computational effort: Mahalanobis distance between y_t and A_t is minimized by gradient descent method. When b_t is got, the gradient is updated by numerical differences. Particle number is set as a monotonically increasing function of error after convergence.

If only standard resampling is applied, the sample with high weight, which may be wrong state, will be replicated many times; the sample with low weight, which may be true state, will hardly be chosen. So PERM sampling[23] is added before standard resampling here: for the sample whose weight π_t^j is lower than π^- , it is accepted with probability a , and the weight after acceptance is $\pi_t^j = \pi_t^j / a$; for the sample whose π_t^j is higher than π^+ , it is resampled K times, and the weight becomes $\pi_t^j = \pi_t^j / K$; otherwise the sample is reserved invariantly. π^-, π^+, a, K are adjusted online: if the tracking quality index (will be stated in (5)) is small, the influence of PERM sampling is weakened, i.e., π^+, a is increased, π^-, K is decreased. Otherwise, π^+, a is decreased, π^-, K is increased.

Estimation of Eyes Blink Amplitude. For the vertices of 2D triangle patches in GNFI, we let the coordinates of those in lower border of upper lid equal to the coordinates of those in upper border of lower lid, so the eyes of GNFI are kept in a closed situation, as so as for the means of OAMs. Therefore, when the current frame opens eyes, eyes blink amplitude is estimated by the consistency between measurements and OAMs.

Coping with Occlusion. Occlusion is judged by the visible situation of 3D triangle patch of 3D facial model as follows: different color values are set to each patch, then the patch is projected on screen under b_t , and the cached color value is read at barycentric coordinate of projected triangle patch. If the read value is same as given value, the patch is not occluded, otherwise, it is occluded.

For j th pixel position in k th 2D triangle patch(fig. 2(c)) of GNFI, $Tr_t^{(k,j)}$ is color value, $Tr\mu_{s,t-1}^{(k,j)}, Tr\sigma_{s,t-1}^{(k,j)}$ are values of mean and variance of A_{t-1} in it, then if k th 3D triangle patch is occluded, $Tr_t^{(k,j)}$ is estimated as:

$$\begin{cases} Tr_t^{(k,j)} = Tr_{t-1}^{(k,j)} + Tr\mu_{s,t-1}^{(k,j)} + Tr\sigma_{s,t-1}^{(k,j)} & Tr_{t-1}^{(k,j)} \geq Tr\mu_{s,t-1}^{(k,j)} \\ Tr_t^{(k,j)} = Tr_{t-1}^{(k,j)} + Tr\mu_{s,t-1}^{(k,j)} - Tr\sigma_{s,t-1}^{(k,j)} & Tr_{t-1}^{(k,j)} < Tr\mu_{s,t-1}^{(k,j)} \end{cases} \quad (4)$$

For Gabor wavelet coefficients, it is similar to above.

Hair Detection from Facial Model Adaptation Result. According to facial model adaptation result, skin is assumed to be present at 3 regions, two are below eyes and one at forehead. The skin color model follows [24]. Hair is present at 3 principle locations adjacent to skin, i.e., the right, middle, left sides of upper face. In initial areas around above locations, skin color model is used to identify non-skin pixels which form seed to model hair color in each area. The seed is iteratively refined by computing the model of color of the areas above current areas, and examining if this color is close to seed. If it is close, the current model is recalculated. The process ends when the color of an area is not close to seed color, thus rough hair detection area is obtained. Finally this area is set as input of Matting algorithm[25] to obtain exact hair detection result.

Updating Textures. The profiles of face and hair are critical to the reality of decoder, thus the textures is updated as follows: when face undergoes significant pose, the image parts whose pixel coordinates are in the region enclosed by outer outline of face model projection and hair detection result are extracted.

3.2 Combining OAM and Cylinder Head Model(CHM)

As simple geometric heads track well the global head motion under large pose variation, CHM is used as it is more suitable for approximating the shape of face than ellipsoid model. Moreover, OAM is suitable for local motion tracking, but works only when pose is deviated not too much from frontal view. So we combine OAM and CHM to get following advantages: 1)the pose angle of successful tracking extends as CHM can provide accurate initial parameters; 2)the result of OAM fitting is used to initialize CHM and enables it to recognize expressions, etc.

The motion of CHM can be parameterized by h in (2). Assuming the intensity of pixel has consistence between adjacent frames, The Δh between t and $t + 1$ can be obtained as: $\Delta h = -\left(\sum_{\Omega} (I_{sg} F_h)^T (I_{sg} F_h)\right)^{-1} \sum_{\Omega} \left(I_{tg} (I_{sg} F_h)^T\right)$, where

$F = \begin{bmatrix} x - y\theta_z + z\theta_y + t_x \\ x\theta_z + y - z\theta_x + t_y \end{bmatrix} \times \frac{f}{-x\theta_y + y\theta_x + z + t_z}$, f is focal length, Ω is the region of I_t whose corresponding pixel of point $x_t = [x, y, z]$ at $t + 1$ is visible, I_{sg} and I_{tg} are the spatial and temporal image gradient, and $F_h = \partial F / \partial h |_{\Delta h=0}$.

Once CHM is fitted to current frame, 3D rotation and translation parameters are obtained and set as the initial global motion parameters of OAM. Once OAM is fitted to current frame, 3D rotation and translation parameters are obtained and set as the initial parameters of CHM in next frame, and the region occupied by the projection of CANDIDE3 is took as the initial template image of CHM.

3.3 Stitching 3D Face, 3D Hair with 2D Background

The current frame is partitioned into squares as H.264 does firstly, then facial motion is tracked. For squares out of the projections of facial model and hair

model, H.264 coding is applied. For square in the border of background and above projections, it is still encoded using H.264, and if the pixel is in the background, bool information of it is set true, otherwise, it is set false.

4 Decoder

Facial animation is performed combining parameterized-based model and muscular-based model. The former adopts RBF, the latter adopts Waters model[18].

The vertices of 3D model are divided to main features, secondary features and non-features. Main features are vertices related to MPEG-4 FDP, secondary features are vertices on muscles around main features, non-features are other vertices. As facial motion expresses different characteristic in different parts, several action areas are defined in 3D facial model, i.e., left/right brows, left/right eyelids, upper/lower lips, left/right lip corners, chin, upper/lower teeth, tongue and rest. Moreover, vertices and facial animation parameters are allocated into these areas according to the influence scopes of action areas. Consequently, in an action area, the displacements of main features are equal to the values of facial animation parameters, the displacements of secondary features are calculated by Waters model, the displacements of non-features are calculated by the fast and effective RBF. Compared to [18] which all vertices in influence scope are calculated by Waters model, it reduces computation greatly while sacrifices tiny reality.

The hair model is adapted as follows: several 2D feature points in hair detection result and corresponding vertices in 3D hair model[21](fig. 1(c)) are selected, then 3D model is adjusted to a individual model using RBF so that the projection of 3D model in image plane is identical to hair detection result.

The texture maps are updated as follows: firstly, the profiles transmitted from encoder are transformed by the minus between global motion parameters of current frame and that of first frame, secondly, the opposite profiles are obtained by exploiting face symmetry, thirdly, they are merged to the texture in first frame, then the border after mergence is smoothed by mean filter, finally, the panorama of face and hair is obtained for texture mapping.

3D face, 3D hair and 2D background are stitched as follows: the current frame is partitioned into squares as H.264 does firstly, then facial animation is produced. For squares out of the projection of 3D facial model and hair model, H.264 decoding is applied. For square in the border of background and above projections, it is still decoded using H.264, but only those pixels out of above projections are displayed according to the bool information from encoder.

5 Experiments

Configuration is: AMD Athlon(tm)II X4 640 3.01G, memory 2G, NVIDIA GT200.

5.1 Encoder/Decoder

For Carphone in MPEG-4, the results of facial model adaption and facial motion tracking are showed as fig. 3. The (c) and (d) are the results after updating texture maps. These results look quite like the human in Carphone. Accurate tracking is obtained even in the presence of perturbing factors including occlusions, etc. Especially, eyes blink amplitude is obtained in rightmost image. The tracking process can be downloaded from the URL¹.

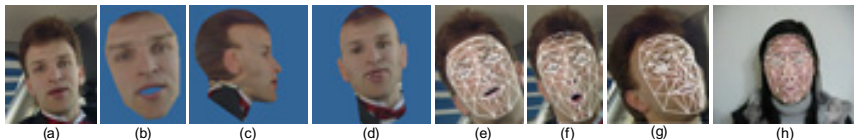


Fig. 3. (a) The first frame. (b) Adaptation result of (a) in encoder. (c) Right profile of adaptation result in decoder. (d) Front of (c). (e)-(h) Facial motion tracking results.

OAM+CHM vs OAM is evaluated by the database[5] that contains global head motion videos and ground truth, and is compared with two measures:

The `tracking_rate` measures the percentage of images which are tracked successfully, where the index in (5) is smaller than a threshold. The `pose_coverage` measures the range of pose angles where the face is successfully tracked.

As fig. 4 and table 1 shows, OAM+CHM-based tracker is more tracking accurate and pose robust than OAM-based tracker under large pose variation.

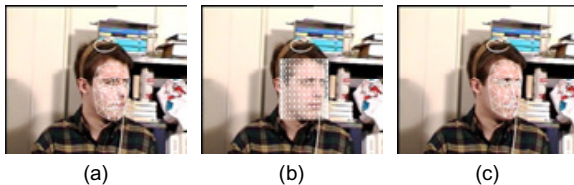


Fig. 4. Facial motion tracking results by OAM(left), CHM(middle) and OAM+CHM(right)

An index is defined to evaluate the tracking quality:

$$Q_T = \sum_{i=1}^N \sum_{j=1}^{M_i} \text{abs} \left(y_{syn}^{(i,j)} - y_{org}^{(i,j)} \right) / (N \cdot M_i) \quad (5)$$

¹ http://staff.ustc.edu.cn/~harryjun/links/facial_tracking.wmv

Table 1. Comparison of the tracking performances

	Tracking rate	Pitch coverage	Yaw coverage	Roll coverage
OAM	0.67	(-18.1, 3.4)	(-35.9, 35.8)	(-18.7, 16.3)
OAM+CHM	0.83	(-19.3, 4.9)	(-40.5, 41.7)	(-18.7, 16.3)

N is total frame number of input video, M_i is pixel number in facial region in i th frame of input video, $y_{org}^{(i,j)}$ is j th pixel color value in that region, $y_{syn}^{(i,j)}$ is j th pixel color value in synthesized image generated by CANDIDE3, i th frame in input video, and $b. abs()$ is the function of absolute value.

As one of state-of-the-art tracker based on OAM, [10] and our system are tested on 13 MPEG-4 testing videos, 110 videos in Cohn-Kanade database[26], and 78 captured videos. From table 2, we can see the superiority of our system.

Table 2. The tracking evaluation accuracy of our system and [10]

	Mean elapsed time per frame	Mean tracking quality
Our system	0.09s	3.79
[10]	0.076s	5.4

The ground truth of facial motion can be got indirectly, i.e., use computer graphic technique to render face image in given value, then facial motion is estimated on rendered image for evaluation. We have synthesized 1734 face images of 48 subjects under various poses and lighting. As table 3 shows, our system with multi-measurements is superior to [10] and that with single measurement.

Table 3. The tracking evaluation accuracy of our system and [10]

Angle	Mean error(Roll)	Mean error(Pitch)	Mean error(Row)
multi-measurements	2.38	1.75	1.69
single measurement	2.73	1.94	1.86
[10]	2.71	1.98	1.85

As fig. 5 shows, according to hair detection result from encoder, an individual 3D hair model is obtained in decoder.

As fig. 6 shows, the profiles of face and hair are obtained based on facial motion tracking and hair detection in encoder and the panorama consisting of face and hair is obtained for texture mapping in decoder.

In channel, textures are compressed by JPEG, facial motion parameters are compressed by predictive coding, and scalar uniform quantization is used.

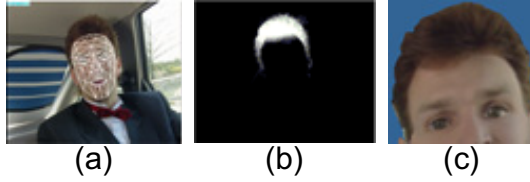


Fig. 5. (a) Facial motion tracking result. (b) The hair detection result. (c) Front of hair model adaptation result.

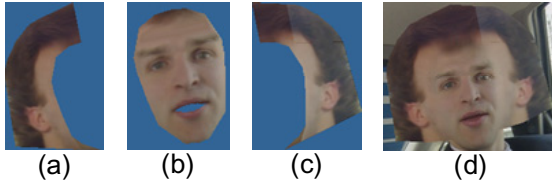


Fig. 6. (a) Left updating texture. (b) The texture in first frame. (c) Right updating texture. (d) The panorama.

Fig. 7 is decoding results of Carphone. An index is defined to evaluate decoding quality: $\sum_{i=1}^N \sum_{j=1}^M \text{abs} \left(y_{ani}^{(i,j)} - y_{org}^{(i,j)} \right) / (N \cdot M)$, where M is pixel number in each frame, $y_{ani}^{(i,j)}$ is j th pixel value in i th frame of decoded video.

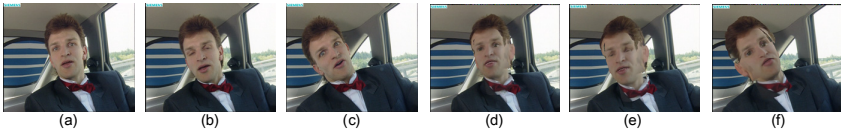


Fig. 7. (a), (b), (c) are input video, (d), (e), (f) are decoded video

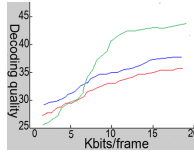
As table 4 shows, facial animation in our system reduces computation greatly while sacrifices tiny reality compared to Waters model.

5.2 Objective and Subjective Evaluation

For objective evaluation, our system, MBC-Face system[20] and H.264 are tested on above videos as fig. 8 shows. Fig. 8 is mean rate/distortion in terms of bit-rate vs PSNR, which is defined as $10 \cdot \text{Log}_{10}^{(255^2/Q_Y)}$. Our system is superior to H.264 in rate/distortion at ultra-low bit-rate, i.e., below 7.2kbit/s, and our system is superior to [20] in rate/distortion at all bit-rates.

Table 4. The comparison of our system and Waters model

	Mean elapsed time per frame	Mean decoding quality
Our system	0.023s	3.9
Waters model	0.07s	3.8

**Fig. 8.** Mean rate/distortion of our system(blue line), [20](red line) and H.264(green line)

For subjective evaluation, our system, [20] and H.264 are performed at 3,5,7 kbit/s, then participants compare avatar in decoded videos with human being in input videos, and complete the questionnaire. Table 5 is the mean scores, and the range is 0-10. Our system obtains the highest scores, indicating that it has more suitability of subjective face identification at ultra-low bit-rate.

Table 5. The mean score of our system, [20] and H.264 in evaluation

Construct	Question	Our system	[20]	H.264
Expressiveness	1. I recognized avatars expressions	7.87	6.73	4.12
	2. Avatars movements looked natural	8.06	7.01	5.54
Appearance	1. I liked avatars appearance	7.83	6.48	4.36
	2. The avatar looks like a human being	8.12	7.47	5.59

6 Conclusion

In this paper, a 2D/3D mixed facial coding/decoding system is proposed. The objective experiment proved it has better rate/distortion performance at ultra-low bit-rate. The between subjects experiment indicated the suitability of subjective face identification at ultra-low bit-rate. The further research directions are: an animation framework including bone, muscle, skin will be established. Facial motion parameters tracked will be used to recognize facial expression.

Acknowledgements. This paper is supported by the Fundamental Research Funds for the Central Universities of China (No. WK2100100009), NSFC (No.61175033), NSFY (No.BJ2100100018) and STP (No.11010202192) of Anhui.

References

1. Liu, Y.C., et al.: A virtual teleconferencing system based on face detection and 3d animation in a low bandwidth environment. *JIST* 20(4), 323–332 (2010)
2. Wang, M., Hong, R., et al.: Movie2comics: Towards a lively video content presentation. *IEEE Trans. on Multimedia* 14(3), 858–870 (2012)
3. Zhang, W., Wang, Q., Tang, X.: Real Time Feature Based 3-D Deformable Face Tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part II*. LNCS, vol. 5303, pp. 720–732. Springer, Heidelberg (2008)
4. Wang, Q., et al.: Real-time bayesian 3-d pose tracking. *TCSVT* 16(12), 1533–1541 (2006)
5. Cascia, M.L., et al.: Fast, reliable head tracking under varying illumination: an approach based on registration of texture mapped 3d models. *TPAMI* 22(4), 322–336 (2000)
6. Ahlberg, J.: Model-based coding: Extraction, coding, and evaluation of face model parameters. PhD thesis, Linköping University, Sweden (2002)
7. Matthews, I., et al.: 2d vs. 3d deformable face models: Representational power, construction, and real-time fitting. *IJCV* 75(1), 93–113 (2007)
8. Jepson, A.D., et al.: Robust online appearance models for visual tracking. *TPAMI* 25(10), 1296–1311 (2003)
9. Lui, Y.M., et al.: Adaptive appearance model and condensation algorithm for robust face tracking. *TSMC* 40(3), 437–448 (2010)
10. Dornaika, F., Davoine, F.: Simultaneous facial action tracking and expression recognition in the presence of head motion. *IJCV* 76(3), 257–281 (2008)
11. Zhou, S., et al.: Visual tracking and recognition using appearance-adaptive models in particle filters. *TIP* 13(11), 1491–1506 (2004)
12. Wen, Z., Huang, T.S.: Capturing subtle facial motions in 3d face tracking. In: *ICCV* (2003)
13. Xiao, J., et al.: Robust full-motion recovery of head by dynamic templates and registration techniques. *IJIST* 13, 85–94 (2003)
14. Fidaleo, D., Medioni, G.G., Fua, P., Lepetit, V.: An Investigation of Model Bias in 3D Face Tracking. In: Zhao, W., Gong, S., Tang, X. (eds.) *AMFG 2005*. LNCS, vol. 3723, pp. 125–139. Springer, Heidelberg (2005)
15. Parke, F.I., Waters, K.: *Computer facial animation*. Wellesley (1996)
16. Blanz, V., et al.: Reanimating faces in images and video. *Computer Graph* (2003)
17. Wang, W.M., et al.: A physically-based modeling and simulation framework for facial animation. In: *ICIG* (2009)
18. Marcos, S., Bermejo, J.G.G., Zalama, E.: A realistic facial animation suitable for human-robot interfacing. In: *ICIRS* (2008)
19. Ward, K., et al.: A survey on hair modeling: styling, simulation, and rendering. *TVCG* 13(2), 213–234 (2007)
20. Eisert, P., et al.: Model-aided coding: a new approach to incorporate facial animation into motion-compensated video coding. *TCSVT* 10(3), 344–358 (2000)
21. Balci, K., et al.: Xface open source project and smil-agent scripting language for creating and animating embodied conversational agents. In: *ICM* (2007)
22. Kampmann, M.: Automatic 3-d face mode adaption for model-based coding of videophone sequences. *TCSVT* 12(3), 172–182 (2002)

23. Grassberger, P.: The pruned-enriched rosenbluth method: simulations of theta polymers of chain length up to 1, 000, 000. *Physical Review E* 56, 3682–3693 (1997)
24. Yacoob, Y., Davis, L.S.: Detection and analysis of hair. *TPAMI* 28(7), 1164–1169 (2006)
25. Levin, A., et al.: Spectral matting. In: *CVPR* (2007)
26. Kanade, T., et al.: Comprehensive database for facial expression analysis. In: *International Conference on AFGR* (2000)

Hierarchical Text Detection: From Word Level to Character Level

Yanyun Qu, Weimin Liao, Shen Lu*, and Shaojie Wu

Computer Science Department, Xiamen University, 361005, P.R. China
{quyanyun, liaoweimin0909, hiplryano, wushaojie246}@gmail.com

Abstract. Text detection is a challenging task in computer vision. In this paper, we focus on English text detection in a natural scene image. We propose a hierarchical approach for text detection, which unifies the word-level text detection and character-level detection as well as the text spatial layout. In our approach, we firstly use stroke width transformation (SWT) to filter an image in a word level. Secondly, we employ the random forest to select discriminative features of characters and compute the confident values of characters. Finally, we use conditional random field to integrate the discriminative information with the text spatial layout, which separates the text from the background. The proposed approach is implemented on the ICDAR dataset, which is a challenging dataset for text detection, and the experiment results demonstrate that our approach is efficient and effective, and it is superior to the state-of-the-art methods in comprehensive criteria.

Keywords: stroke width transformation, random forest, condition random field, spatial layout, multiple channel of text features.

1 Introduction

Text recognition is one of the important tasks in computer vision. The text in the natural scene image provides significant clues for image understanding. For example, the text in a sign board or the text in a traffic sign can present the indicative information, such as where the road goes or what we should take care of. Therefore, text recognition has the widely applied prospect, and it is critical in content-based image retrieval, assistive navigation, scene understanding and geography annotation, etc. Although, the optical character recognition (OCR) has been successful in the application, it is suitable to deal with the scanned document in which the text is not interfered by the background and is easy to be distinguished from the background. Thus, text recognition in the wild is still a challenging problem.

Text recognition includes two sequential parts, one is text detection and the other is text recognition. The former focuses on verifying if there is any text in an image and where the text is. The latter focuses on recognizing the text given

* Corresponding author.

a word image, that is, it identifies the words or letters, after the text is localized. These two parts are usually studied separately.

This paper focuses on text detection which is the prerequisite for text recognition. Text detection in a wild is a very challenging task, for the text is in a variety of appearances. As shown in Fig. 1, there are many difficulties to localize the text in a natural scene image, and the text appearance is affected by illumination, scale, image resolution, font, viewpoints, spatial layout and background, etc. There are lots of methods to solve this problem. Some are based on rules to detect the text after obtaining the connected regions, and some are based on the texture feature learning. However, those methods are not robust to the variance of the text appearance and their detection performance need to be improved.

In this paper, we aim at text detection by fusing the character features and the word features. We are inspired by the following observation: 1) the text in a natural scene image usually has the similar stroke width, 2) the character usually has the distinctive texture cues compared with other objects, 3) the characters in a word are usually arranged in a uniform way, and the interval between each pair of characters in a word is similar. Thus, the characters in a word are of distinctive spatial layout. We propose a hierarchical approach for text detection. In the first stage, we select the stroke width as a feature to filter a text image. In the second stage, we transform the candidate regions obtained in the first stage into feature maps in multiple channels and use the random forests to compute the confidence value. In the third stage, we unify the discriminative information obtained from the random forest with the spatial layout of a word based on conditional random field (CRF). The contribution of our approach is to unify the word level features and the character level features to localize texts in a natural scene image.

The paper is organized as follows. In Section 2, we briefly discuss the related work. The word-level text detection and the character-level text detection are



Fig. 1. Challenges facing text localization. a) Illumination changes, b) Viewpoint changes, c) Cluttered background, d) Scale changes, e) Different layouts, f) Different font styles.

detailed in Section 3 and Section 4 respectively. And text localization based on CRF are introduced in Section 5. Conclusions are given in the last section.

2 Related Work

Until now, many methods focus on text detection and recognition. Text recognition in the wild is a hot topic recently. Some researchers paid attentions to recognizing the words or letters in a word image which only contains the text [15][18]. Some researchers studied the end-to-end text recognition [14][19], in which a systematic framework from the text detection to text recognition was designed. But the most important thing for text recognition is how to obtain the text region. There are many related methods, which are classified into two categories: rule-based methods and learning-based methods.

The rule-based methods firstly extract connected regions, and then design the rules to distinguish text from the background. Lienhart [3] and Jung [4] used connected component analysis (CCA) to localize the text in images and video frames. Liu [17] proposed a stroke width filter to extract the text structure. Becker, who won the first prize of the ICDAR 2005 competition [5], designed the adaptive threshold and the rules of text detection, further implement the vanishing point to extract text. Lyu [6] proposed a coarse-to-fine video text detection, localization, and extraction method for the multilingual text localization. In order to extract the text, he first implemented edge detection, and then he designed thresholds for edge map, local edge density, and horizontal and vertical projection, etc. Shivakumara [7] designed heuristic rules to classify low contrast and high contrast video images combined with edge filter. Hua [8] used a corner-based method to automatically localize text in video frames. Thillou [9] extracted color texts in natural scenes with selective metric-based clustering. Ye [10] proposed a novel coarse-to-fine algorithm based on multi-scale wavelet features to locate text lines in complex background. Epshtein [12] proposed a stroke width transform to improve the detection of text candidate regions. Liu [13] performed edge detection in four directions and extract features on edge maps for text and non-text classification. The limit of the rule-based methods is that the rules are not robust and they fail to remove the background noise from the text.

Different from the rule-based method, learning-based methods learn the rules from the text training data, thus they are more robust. Chen [11], who won the second prize of the ICDAR2005, made text statistic to select the discriminative features, and use the Adaboost classifier to separate texts from the background. Jung [1] proposed a framework based on SVM to perform accurate text localization. He extracted the text intensity features and constant gradient variant (CGV) features to learn a SVM classifier. In his another paper, Jung [2] designed a stroke filter combined with a SVM classifier to extract text regions. Kim [4] analyzes the textural features in natural scene images using an SVM classifier and locates the text regions by operating CAMSHIFT to group the text regions. The most related work to ours is Wang's method [14]. He proposed a systematic

approach to recognize texts. He first used HOG classifier to compute the confident value of a region, and then he employed the pictorial structure to constrain the text spatial layout. The difference is that his approach supposed that the words which were present in an image were given.

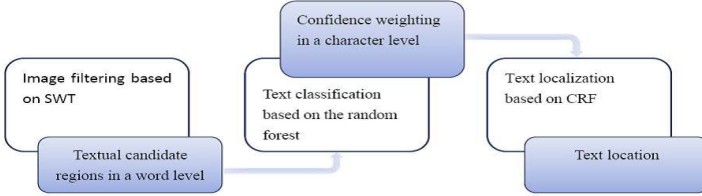


Fig. 2. The framework of our approach

The methods mentioned above are mainly concerned about a single level features, they may cause larger false positive detection rate. Therefore, we propose a hierarchical approach which combines the word-level features and the character-level features for text localization. In a word level, we use the stroke width transform to filter an image; in a character level, we design random forests to select the discriminative features of characters and compute the confident value of the candidate region. Furthermore, we implement CRF to unify the discriminative information with the spatial layout of words to separate the text from the background. The framework of our approach is shown in Fig. 2.

3 Text Detection in a Word Level

As we observe, the English text in natural scene usually is arranged in a uniform way, that is, the characters in the text have the similar stroke width, which is shown in Fig.3 a). Therefore, in a word level, we employ the SWT to filter an image. Each pixel value of the transformed image is equal to its stroke width, and the size of the SWT image is equal to its original image.

In detail, an image is firstly processed by the Canny edge operator, and then the gradient orientations are computed. If a pixel p locates at the edge of a stroke, its gradient orientation d_p is perpendicular to the stroke orientation. And the radial $r = p + n * d_p$ ($w_{max} > n > 0$) would intersect at the other edge of the stroke, where w_{max} is the maximum value of the stroke width and the intersection point at the other edge of the stroke was defined as q . The gradient orientations of the two points p and q are inverse, which satisfies $d_q = -d_p \pm \tau$, where τ is the error which is not greater than a threshold. The pixel values of points in the line between the points p and q are set $\| \overrightarrow{p - q} \|$. If a point p does not find its corresponding point q at the other edge of the stroke, the radial r will be deleted. In our experiment, we set $w_{max} < W_{img}/10$, that is, the maximum

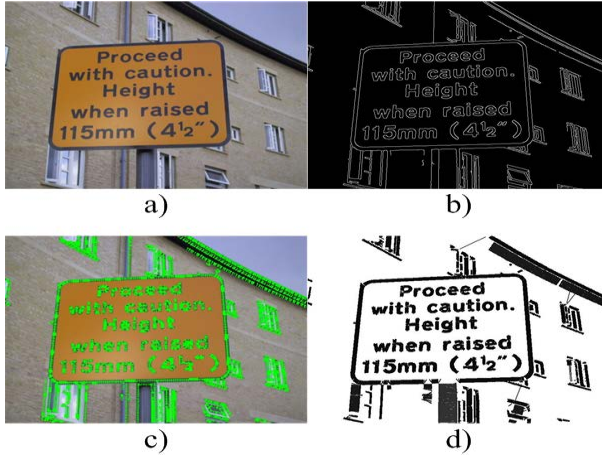


Fig. 3. The middle results of the SWT. a) The original image, b) The edge image, c) The gradient orientation image, d) The SWT image, where the darker the pixel is, the smaller the SWT value is.

of stroke width is not greater than one tenth of the width of the image, and $\tau = \pi/6$. Fig. 3 shows the intermediate results of the SWT.

Furthermore, we analyze the connected components and use the mathematical morphologic filter(MMF) to filter the non-text regions. The results of the SWT are shown in Fig. 4.



Fig. 4. The SWT results combined with CCA and MMF. The green boxes are the textual candidate regions, and the red regions are filtered by CCA and MMF.

4 Text Detection in a Character Level

4.1 Multiple Channels of Text Features

In this subsection, we introduce the proposed text detection in a character level which uses text multiple features. There are four types of text features which are

widely used in text localization: normalized gray intensity (N-gray), color cues, constant gradient variance (CGV), and histogram of gradient (HOG).

N-gray is proved to be useful for text detection by Kim and Jung [1] [4]. They empirically demonstrated that N-gray is better than wavelet feature as well as the histogram descriptor in terms of text description, and they also find that N-gray can deal with the low contrast document. The N-gray feature of a pixel is defined as $Nf(s) = (f(s) - V_{min}) / (V_{max} - V_{min}) * 255$, where s is the center point of a sub-window, $f(s)$ is the original gray intensity of pixel s , V_{min} and V_{max} represent the minimum intensity value and the maximum intensity value respectively in this sub-window. The N-gray is the normalized result which transforms the gray value to the interval $[0, 255]$.



Fig. 5. Color cues based on Lab color space

The second type of text features are color cues. In order to avoid the illumination inference, we adopt the Lab color space. Each sliding window contains three color channels of L , a and b , which is shown in Fig. 5.

Motivated by the effectiveness of CGV for text detection in the work [10], we adopt CGV as the third feature, which is shown in Fig.6. The CGV is used to normalize the contrast at a given pixel. It is defined as

$$CGV(S) = (g(s) - LM(s)) \sqrt{\frac{GV}{LV(s)}}. \quad (1)$$

where $LM(s) = \frac{1}{|w_s|} \sum_{s_i \in w_s} g(s_i)$ and $LV(s) = \frac{1}{|w_s|} \sum_{s_i \in w_s} (g(s_i) - LM(s))^2$. In this equation, $g(s)$ denotes the gradient value at pixel s and w_s denotes the neighborhood of pixel s . $LM(s)$ denotes the local mean of the gradient of the neighborhood of pixel s , $LV(S)$ denotes the local gradient variance, GV denotes the global gradient variance computed over the whole image window, and $CGV(s)$ denotes the feature value at pixel s .

The fourth type of text features is the Histogram of Oriented Gradients (HOG) proposed by Dalal [16]. Motivated by its success in pedestrian detection, we use it to describe the text. HOG is a 3D histogram. In our approach, we normalize a character image to 48×48 , and the normalized image is tiled with the grid of overlapping blocks whose size is 5×5 . Each block contains a grid of spatial cells.

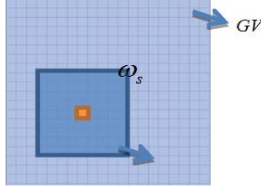


Fig. 6. Constant Gradient Variance

We slide the cell every pixels. For each cell, the magnitudes of image gradients weight orientation histograms with 9 bins, and the histogram in each cell is normalized. Finally, a character image is represent by a feature vector with $48 \times 48 \times 9$ dimensions.

Therefore, the four types of features form the multiple channels of features. Take the HOG feature as an example, it forms 9 channels, each of which has the same size as the original normalized image.

4.2 Confident Weighting Based on Random Forest

Considering that it is time consuming to directly use all channels of features, we implement the random forest to select the discriminative features from the four types of features, because of the success of the random forests in multiple feature selection and multi-class classification.

A random forest classifier is a popular classification method. It is an ensemble classifier that consists of many decision trees, and the label of a query image is determined by the majority of the votes which is shown in Fig. 7. The random forest is made up of several decision trees. Given the training set T , the data pair in T is denoted as $(x_1, x_2, \dots, x_n, y_i) \in T$, where x_i is the i th dimension of the feature vector, y_i is the label. Each tree is constructed in the same way. In detail, a tree is constructed by recursively splitting T into two subsets T_{left} and T_{right} . The chosen feature x_i can best split the training data set T , that is, it can result in the largest expected information gain of the node categories, which is calculated as:

$$Gain(x_i) = Entrop(T) - \frac{T_{left}}{T} \times Entrop(T_{left}) - \frac{T_{right}}{T} \times Entrop(T_{right}) \quad (2)$$

where T_{left}, T_{right} are subsets of T and are splited by x_i . The information entropy is defined as:

$$Entrop(\psi) = - \sum_{i=1}^{|C|} freq(C_i, \psi) \times \log_2(freq(C_i, \psi)) \quad (3)$$

where C_i denotes the set of classes, $freq(C_i, \psi)$ denotes the occurrence frequency of C_i in set ψ . Each branch node in a tree has a split test associated with it, and one or the other child is chosen based on the result of the test.

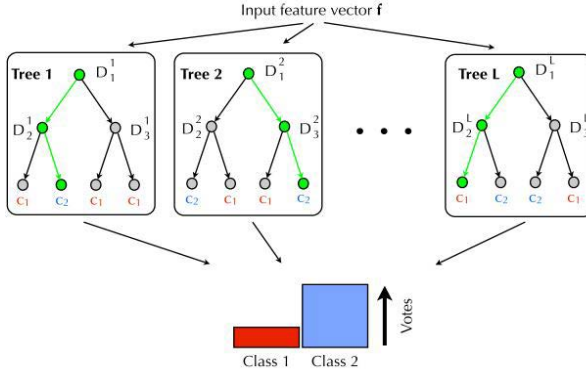


Fig. 7. The random forest classifier

In the training stage, we collect the character images from the ICDAR 2003. We also randomly sample the background as the negative samples. Thus, we construct a training set which contains 6110 positive samples and 4196 negative samples. Some training samples are shown in Fig.8. We train the random forest classifier on the training set.

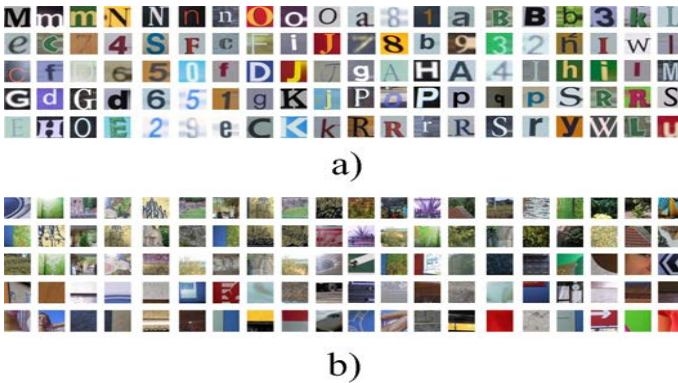


Fig. 8. Some samples in the training set. a) Some positive samples, b) Some negative samples.

Now we implement the random forest to select discriminative features and compute the confident values of the candidate regions. The flowchart of the character-level text detection is shown in Fig. 9. We firstly implement the sliding windows scheme on the candidate regions obtained from SWT, and then we compute the confident weights based on the learned random forest classifier. After that, we obtain a confident map where each pixel value is equal to its confident value.

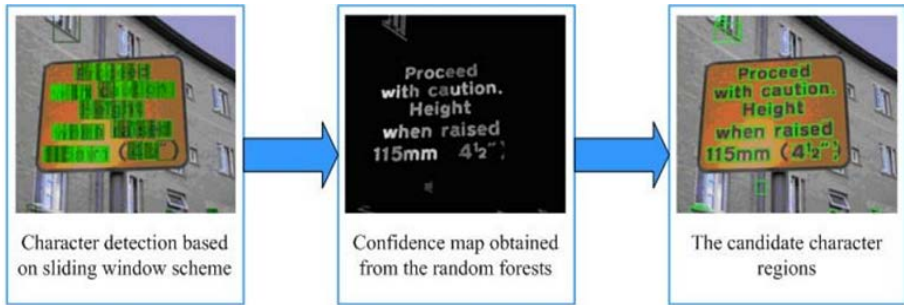


Fig. 9. The flowchart of the character level text localization

5 Text Localization Based on CRF

In this section, we unify the discriminative information of texts with the text spatial layout based on CRF. As we observe, the spatial layout in a word or in the text is characteristic for text detection. If we see a character, we always expect there are some characters in its left neighbor or its right neighbor. It is true in a statistic view point.

We construct the graph model $G = (V, E)$, where V is the vertex set, and the E is the edge set. If the pairwise candidate characters satisfy the following condition,

$$\text{dist}(\chi_i, \chi_j) < 2 * \min(\max(w_i, h_i), \max(w_j, h_j))$$

we add an edge between these two letters, where $\text{dist}(\cdot, \cdot)$ is the distance between the centers of two characters, w and h is the width and the height of a character respectively. In CRF, the unary potential is defined as the confident value obtained from the random forest, and the binary potential is defined as the covariance distance between the query letter location and the expectation location. The parameters of the covariance matrix are obtained by training on the ICDAR2003 dataset.

6 Experimental Results

In this section, we implement our approach on the ICDAR2005 dataset, which is similar to ICDAR2003. It contains color images of all kinds of scenes taken by digital cameras with a range of resolution and other settings. The images include household objects, road signs, board signs, bill-boards and posters. The ICDAR2005 dataset contains three parts: sample, trail and competition. We only use trail set in our approach. The trail set contains 258 training images and 251 testing images. The visual results are shown in Fig. 10, and the results of our approach are signed by green rectangles. Fig. 10 demonstrates that our approach is effective.



Fig. 10. Some text localization results obtained by our approach

We also evaluate our approach by the following criteria: the recall, the precision, the F-score, and the running time in seconds according to ICDAR2005 [5]. We compute the average time processing an image in the testing set. We compare our approach with 9 competition methods which are from ICDAR 2003 competition and ICDAR2005 competition [5] for text localization. As shown in Table 1, the detection precision of our approach is 65%, which is the highest precision; the recall rate is 59% , which is ranked in the third; the F-score is 62% which ranks the top with Hinnerk’s method; the averaged running time is 3.87 seconds which is at a moderate level. To sum up, our approach achieves the best comprehensive performance in text localization.

Table 1. The performance comparison of text localization

Methods	Precision	Recall	F-Score	Time
Our Methods	0.65	0.59	0.62	3.87
Hinnerk	0.62	0.67	0.62	14.4
Alex Chen	0.60	0.60	0.58	0.35
Qiang Zhu	0.33	0.40	0.33	1.6
Jisoo Kim	0.22	0.28	0.22	2.2
Nobuo Ezaki	0.18	0.36	0.22	2.8
Ashida	0.55	0.46	0.50	8.7
HWDavid	0.44	0.46	0.45	0.3
Wolf	0.30	0.44	0.35	17.0
Todoran	0.19	0.18	0.18	0.3

7 Conclusions

In this paper, we propose a hierarchical approach for text localization in a natural scene image. The advantage of our approach is to combine the word-level features with the character-level features, as well as the spatial layout of the text. We compare our approach with 9 competition methods which are from the competitions of ICDAR2005 and ICDAR2003 in terms of precision, recall, F-scores and running-time. We implement our approach on ICDAR2005, and the experimental results demonstrate that our approach achieves the best precision and F-score. And our approach are also superior in the comprehensive performance.

Acknowledgments. This research work was supported by the Fundamental Research Funds for the Central Universities (2010121067) and National Defence Basic Scientific Research program of China(B1420****55).

References

1. Jung, C., Liu, Q., Kim, J.: Accurate text localization in images based on SVM output scores. *Image and Vision Computing* 27, 1295–1301 (2009)
2. Jung, C., Liu, Q., Kim, J.: A stroke filter and its application to text localization. *Pattern Recognition Letters* 30, 114–122 (2009)
3. Lienhart, R., Effelsberg, W.: Automatic text segmentation and text recognition for video indexing, TR-98-009, University of Mannheim (1998)
4. Jung, K., Kim, J.: Texture-based approach for text detection in images using support vector machines and continuously adaptive mean shift algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25(12), 1631–1639 (2003)
5. Lucas, S.M.: ICDAR 2005 text locating competition results. In: *Proceedings of the Eighth International Conference on Document Analysis and Recognition*, pp. 80–84 (2005)

6. Lyu, M.R., Song, J., Cai, M.: A comprehensive method for multilingual video text detection, localization, and extraction. In: *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 243–255 (February 2005)
7. Shivakumara, P., Huang, W., Phan, T.Q., Tan, C.L.: Accurate video text detection through classification of low and high contrast images. *Pattern Recognition* 43, 2165–2185 (2010)
8. Hua, X.-S., Chen, X.-R., Wenxin, L., Zhang, H.-J.: Automatic location of text in video frames. In: *Proceedings of the 2001 ACM Workshops on Multimedia: Multimedia Information Retrieval*, October 05 (2001)
9. Thillou, C.M., Gosselin, B.: Color text extraction with selective metric-based clustering. *Computer Vision and Image Understanding* 107, 97–107 (2007)
10. Ye, Q., Huang, Q., Gao, W., Zhao, D.: Fast and robust text detection in images and video frames. *Image and Vision Computing* 23(6), 565–576 (2005)
11. Chen, X., Yuille, A.L.: A time efficient cascade for real-time object detection: with applications for the visually impaired. In: *Proceedings of the CVAVI 2005, IEEE Conference on Computer Vision and Pattern Recognition Workshop* (2005)
12. Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 13–18, pp. 2963–2970 (2010)
13. Liu, C., Wang, C., Dai, R.: Text Detection in Images Based on Unsupervised Classification of Edge-based Features. In: *Eighth International Conference on Document Analysis and Recognition (ICDAR 2005)*, pp. 610–614 (2005)
14. Wang, K., Babenko, B., Belongie, S.: End-to-end scene text recognition. In: *2011 IEEE International Conference on Computer Vision (ICCV)*, November 6–13, pp. 1457–1464 (2011)
15. Wang, K., Belongie, S.: Word Spotting in the Wild. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010, Part I. LNCS*, vol. 6311, pp. 591–604. Springer, Heidelberg (2010)
16. Dalal, N.: Finding people in images and videos, France: the French National Institute for Research in Computer Science and Control. In: *INRIA* (2006)
17. Liu, Q., Jung, C., Moon, Y.: Text Segmentation based on Stroke Filter. In: *Proceedings of International Conference on Multimedia*, pp. 129–132 (2006)
18. Mishra, A., Alahari, K., Jawahar, C.V.: Top-Down and Bottom-Up Cues for Scene Text Recognition. In: *CVPR* (2012)
19. Neumann, L., Matas, J.: A Method for Text Localization and Recognition in Real-world Images. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) *ACCV 2010, Part III. LNCS*, vol. 6494, pp. 770–783. Springer, Heidelberg (2011)

Building a Large Scale Test Collection for Effective Benchmarking of Mobile Landmark Search

Zhiyong Cheng¹, Jing Ren¹, Jialie Shen¹, and Haiyan Miao²

¹ School of Information Systems, Singapore Management University
{zy.cheng.2011,jing.ren.2012,jlshen}@smu.edu.sg

² Institute of High Performance Computing, A*STAR, Singapore
miaohy@ihpc.a-star.edu.sg

Abstract. Studying and analyzing system performance is one of the fundamental factors for the related technological advancement in image retrieval. In this paper, we report the construction of a large scale test collection for facilitating robust performance evaluation of mobile landmark image search. Totally, the test collection consists of (1) 355,141 images about 128 landmarks in five cities over 3 continents from Flickr; (2) different kinds of textual features for each image, including surrounding text (e.g. tags), contextual data (e.g. geo-location and upload time), and metadata (e.g. uploader and EXIF); and (3) six types of low-level visual features. For the task of landmark image retrieval evaluation, we also conduct a series of baseline experimental studies on the search performance over different visual queries, which represent different views of a landmark.

Keywords: Mobile Landmark Search, Performance Evaluation.

1 Introduction

With the recent explosion of consumer generated images and the fast growth of photo sharing websites, research interest in social image search has been growing in information retrieval, machine learning and multimedia systems. Many different algorithms or systems have recently been developed to support automatic retrieval or visualization of landmark [6,5,18,17]. In particular, techniques for large scale mobile landmark image search are becoming more and more important due to a wide range of real applications [8]. As a result, lots of efforts have been devoted to improve the search systems' performance from different aspects (e.g. retrieval accuracy [9] and diversification of the displaying results [14,2]).

Studying and analyzing system performance is one of the fundamental factors to enable the technology advancement in landmark search. However, very limited work has been done on developing benchmarking dataset to compare and evaluate relative techniques and systems. While the importance of the issue has been recognized in the multimedia retrieval community and a few test

collections have been published recently, they generally suffer from one or multiple weaknesses as follows: small scale, unclear search task definition, diversified views of landmarks and limited availability. The problems can be particularly severe when the researchers try to do cross-method comparison. Due to the lack of quality collections, they are forced to construct their own datasets by leveraging online public resources, such as Flickr¹ and Google image [9,14,2,10,4,16]. This can easily lead to very expensive and tedious data collection process. More importantly, the use of self-constructed datasets makes it hard for other scholars to repeat the experimental studies and compare different methods to assess the precise impacts of individual systems.

In principle, the procedure for the performance evaluation can be generally divided into four basic steps: 1) construct a test image collection; 2) generate a set of test images and the related ground truth information; 3) run each test image through a particular landmark search system; and 4) assess performance of the system via an empirical distribution of particular measurement metric (e.g., precision or recall ratio). All four steps are critical for the quality of performance evaluation. In this paper, we focus on the study of test collection. We argue that the following three criteria are basic guidelines for achieving reliable and accurate assessment results in developing testbed: Since images in many real photo sharing websites have scaled to billions over the last few years, the first guideline is to keep test collection’s scale big enough. The second guideline is to good visual coverage of different geographic locations. We argue that when more partial views about the same landmark are included, evaluation process can have more completed and thus more accurate results can be expected. The last guideline is that search tasks for evaluation and associated ground-truth information should be clearly defined.

Motivated by the discussion above, we create a large scale benchmarking dataset to support effective and robust comparison of landmark search system performance. Totally, the test collection consists of 355,141 images about 128 landmarks in five cities over three continents from Flickr. Besides, six different visual features are extracted as signature for each image. For each landmark, a wide range of partial views have been considered to gain comprehensive visual coverage. Moreover, we give a clear definition of different search tasks and ground truth information. Using them, a set of empirical studies have been carried out to investigate the search accuracy and efficiency of content-based search methods, using different views of one landmark as queries.

2 Test Collection Construction

The construction of the landmark image dataset starts from selecting a set of international cities that contain various popular landmarks. At this stage, five cities from 3 continents are considered, including *Beijing*, *Hong Kong*, *Singapore*, *London*, and *New York*. Well-known landmarks of each city are selected from the

¹ <https://www.flickr.com/>

landmark lists published in Wikipedia² and Wikitravel³ (we also refer to other online tour guide web pages). Altogether, 128 landmarks are identified in the five cities. The number of landmarks in each city can be found in Table 2. Table 5 shows the details of landmarks in Singapore. The details about the whole dataset can be found in [1].

2.1 Dataset Downloaded

The images of each landmark are collected from Flickr. For the landmark with an unique name, its name is used as the keywords to search images. While for the landmark whose name also corresponds to other landmarks in different cities, the name of corresponding city is also included in the search keywords. For example, we use “*city hall, singapore*” and “*city hall, new york*” to search the images of City Hall in Singapore and New York, respectively. We retrieve the most relevant images based on the tag-based method provided by Flickr’s public API⁴. This method requires the returned images must contain the query terms in their tags, and the returned images are sorted in descending order based on relevance. The top 4000 images in the returned list are taken. Notice that not all images in the list can be successfully downloaded. Besides, some landmarks have less than 4000 images tagged with their names in Flickr. Thus, the number of downloaded images for each landmark is 3301 on average before processing. Different kinds of related data associated with each image are also collected. We categorize the associated information into three types: *surrounding text*, *context information* and *metadata*. Details can be found in Table 1. The surrounding text includes title, tags, description and comments, which directly represent the semantic features of the image. We consider four different kinds of context information: *taken time* is about when the image was captured; *upload time* refers to the time when the image was uploaded to Flickr; *geo-location* generally is about the location where the image was taken; *contextual URLs* contains the URLs of photostream, sets and pools to which the image belongs. Each image in Flickr has an unique *photo ID*; *uploader ID* refers the ID of the user who contributed the image; *EXIF/TIFF* contains the image metadata, such as the device used to capture the image and parameter-setting of the device at the time of taking the image. The *source page* of the image in Flickr is kept as the backup reference.

Table 1. Related Information Crawled for Each Image

Surrounding Text	Context Information	Metadata
tag	taken time	photo ID
title	upload time	uploader ID
comment	geo-location	source page
description	contextual URL	EXIF/TIFF

² <http://www.wikipedia.org/>

³ http://wikitravel.org/en/Main_Page

⁴ <https://www.flickr.com/services/api/>

2.2 Dataset Statistics

Using the method described above, totally 419,346 images are collected at Flickr’s medium-scale image resolution, which is 500×500 pixels maximum. In the collected images, the most common sizes are “ 500×375 ”, “ 500×333 ”, “ 375×500 ”, and “ 333×500 ”, accounting for 59.04% of the dataset. We remove the image whose length or width is less than 300. Also, if any piece of the related information listed in Table 1 fails to be downloaded, the corresponding image won’t be included to our test collection. Finally, there are 355,141 images left for 128 landmarks. The distribution of image number for landmarks in each city is shown in Table 2. Meanwhile, Table 3 has shown the statistics information of surrounding text of images. Note that more than half of the images do not have any comment, and the minimal number of tags is 1 is because of the used tag-based search method. Fig. 1 shows the distribution of the number of tags per image. In the figure, the number of images is normalized. To assess the quality of the downloaded images, 20,000 images are randomly drawn from the whole dataset and manually evaluated. 735 images are labeled as low-quality, representing 3.675% of the subset. This ratio can be regarded as an indicator of the proportion of low-quality images in the whole dataset.

Table 2. The number of landmarks and the distribution of image number across landmarks in each city

City	Number of landmarks	Distribution of number of images		
		Average	Max	Min
Beijing	25	2874.64	3680	728
London	25	2994.56	3252	1386
Singapore	28	2562.11	3401	589
New York	28	2898.68	3692	741
Hong Kong	22	2523.14	3882	556

Table 3. Statistics of surrounding text in the dataset

	Average	Max	Min
Number of tags	11.24	160	1
Number of keywords in title	4.07	41	0
Number of keywords in description	65.34	14303	0
Number of comments	13.74	2055	0
Number of keywords in comment	4.95	593	1

2.3 Visual Features

For the convenience of utilizing the dataset on the performance evaluations of various applications, we extract and provide a set of effective and widely used visual features for each image. They include,

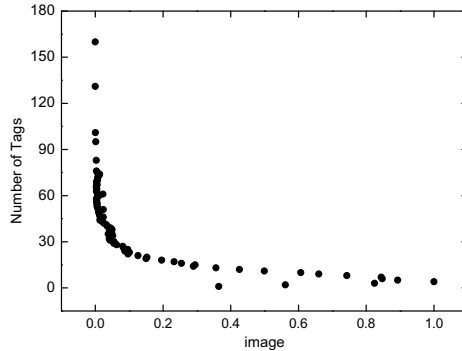


Fig. 1. The number of tags per image

Color Histogram (64D) [15]: The HSV color space is divided into 64 partitions, and the number of pixels within each partition is then counted for computing the histogram bin of the corresponding color.

Color Auto-Correlogram (144D) [7]: The color auto-correlogram describes the global distribution and the spatial correlation of pairs of colors together. We consider the HSV color space with color quantized into 36 bins, and use 4 distance metrics as [7] to compute the auto-correlogram.

Gabor Texture (72D) [12]: Wavelet features are extracted at multiple scales and directions from the images using a Gabor wavelet decomposition. The mean and standard deviation of the filter responses are calculated. We extract Gabor features in six different orientations and six different scales.

Block-Wise Color Moments (255D): Each image is divided into 5×5 grid partitions. For each grid, the first three color moments (*mean*, *variance*, *skewness*) are calculated for each color channel in HSV color space. Each grid region is then characterized by 9 moments, resulting in a 225-dimensional vector for an image.

Edge Histogram (80D) [13]: The edge histogram represents the spatial distribution of five types of directional edges, namely four directional edges and one non-directional edge. Each image is partitioned into 4×4 grid, and each grid is further divided into small square blocks. Five directional edges are extracted from the small blocks. Then the number of five edge types in each grid is counted to define five histogram bins for the corresponding grid.

Bag-of-Visual-Words [11]: 500-D bag-of-visual-words (BoVW) is generated for each landmark. For each image, key-points are detected using difference of Gaussian. Then each key-point is described by a 128 dimensional SIFT descriptor [11]. Finally, the descriptors of each image are vector quantized into a vocabulary of visual words, which are generated by k -mean clustering method.

2.4 Dataset Structure

In order to support fast browsing and exploration of the collection, the dataset is organized in hierarchical structure based on location and landmark categories. Under two general categories - *Natural Attractions* and *Man-made Attractions*, we further define 13 subcategories as: A. Natural Attractions: (A1) beach, (A2) island, (A3) mountain, (A4) nature reserve, (A5) wildlife attractions and (A6) park and garden; B. Man-made Attractions: (B1) buildings and monuments, (B2) distinct small town, (B3) harbor and bay, (B4) historic resort, (B5) museum and gallery, (B6) religious architecture, (B7) shopping and commercial center. Particularly, the landmarks are first separated based on cities and spatial districts; then landmarks in the same district are classified into different subcategories. As an example, Fig. 2 illustrates the hierarchical structure using several landmarks in Singapore. The whole structure of landmarks in Singapore is shown in Table 5.

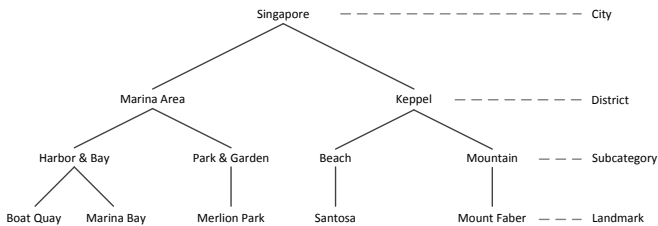


Fig. 2. Hierarchical Structure of the Dataset

3 Applications of Dataset

The dataset is mainly constructed for facilitating the development and evaluation of landmark search systems. Meanwhile, some other related research issues can also be explored using the dataset. In this section, we discuss these related issues and their relations with landmark image search.

Content-Based Landmark Search. While a lot of research efforts have been invested to improve performance of content-based landmark search, many open problems in system development still remain, including (1) various visual appearance of a specific scene under different shot conditions or angles; and (2) landmarks (or part of them) may have similar visual appearance (e.g. church spires), which makes them hard to be discriminated by visual features. The combination of textual features has demonstrated some promising results. However, most of them only consider one type of textual features (e.g. tags). The effectiveness of combining multiple textual features with visual information needs to be further studied. Besides, along with the development of effective search methods, novel indexing methods also need to be designed in order to deal with large-scale datasets.

Landmark Recognition. The ability to recognize the landmark depicted in an image can facilitate the content understanding and geo-location detection in images, and thus can improve the performance of landmark image retrieval. Most of current techniques on this problem rely on accurately labeled data. How to leverage noisy data, such as the dataset crawled from Flickr, to perform accurate landmark recognition is still a challenging task.

Scene Summarization. The scene summarization is to select a set of images that represent the most interesting and important aspects of the landmark with minimal redundancy, such that the viewers can quickly get an accurate impression on the landmark. This technique is very useful in presenting the search results of landmark image retrieval, because users typically tend to know the different representative views of the landmark instead of the redundant images using the same appearance.

4 Empirical Study on Content-Based Landmark Search

We conduct experiments on the dataset to study the performance of landmark search using content-based methods. In particular, we study (1) the search performance of different visual queries which represent different views of a landmark, and (2) the search efficiency of different visual features.

4.1 Experimental Setup

Query Set. We select eight landmarks⁵ in Singapore and take pictures of various views of each landmark. From the taken photos, images which represent the following types of views are selected as queries, including (1) views from different angles (i.e. front views and side views), (2) partial views (i.e. different parts of the landmark), (3) interior views, (4) close-up exterior views, and (5) far-away exterior views. Fig. 3 shows the examples of different views. For each landmark, we select four queries for each type of view. Finally, there are total 156 queries are selected⁶.

Experimental Subset. The targeted landmarks belong to the subcategories of *park and garden*, *buildings and monuments*, *harbor and bay* and *museum and gallery*. We select images of landmarks in those subcategories to construct a challenging distractor subset, as images of landmarks in the same subcategory are more likely to be similar. Images of landmarks in *Singapore*, *New York* and *London* are used. Altogether, 59 landmarks with 164,690 images are included in the subset.

⁵ The selected landmarks include *Armenian Church*, *Cathedral of the Good Shepherd*, *Church of Our Lady of Lourdes*, *Church of Saints Peter and Paul*, *Marina Bay*, *Merlion Park*, *National Museum* and *Raffles City*.

⁶ Because *Merlion Park* is an open area, it does not have queries representing the interior views.



Fig. 3. Different types of views using images of the Cathedral of the Good Shepherd as examples

Visual Features. We use color histogram (CH), color moments (CM), bag-of-visual-words (VW), and two combinations of them (CH + CM and CH + CM + VW) as visual features in the experiments. A vocabulary with 1000 visual words is generated for the subset using the method described in section 2.3. Euclidean distance is used for calculating the similarity score. In the combination of different features, the similarity scores are separately computed and normalized, and then uniformly summed together to obtain the final score.

Ground Truth and Evaluation. The task of landmark image retrieval is to search visual views of the desired landmark, which means that positive results must be or at least contain visual view of the targeted landmark. According to this, the judgement criterion is defined as: if an image contains views of the targeted landmark that are recognizable to viewers, then the image is regarded as positive; otherwise, the image is marked as negative. The top 10 results of each query are assessed by human evaluators. The precision ($P@10$) is used as the evaluation metric.

4.2 Experimental Results and Analysis

All the experiments are conducted on a desktop computer with Intel Core i5 2.80GHz CPU and 4GB memory. In this section, we describe the results and detailedly analyze the performance in terms of accuracy and efficiency.

Accuracy Study. Table 4 shows the search accuracy of different visual queries in each view type. The values in the table are the average $P@10$ over all queries in the same type. From the table, we can see that the performances of color histogram and color moments are comparable, and the feature of visual words performs slightly better. Although the combination of global and local features improves the search accuracy, it is still pretty low. Content-based landmark

Table 4. Average precision of visual queries in different view types. The representations of the acronyms in the table: AV - views from different angles, PV - partial views, IV - interior views, CUV - close-up exterior views, FAV - far-wary exterior views, CH - color histogram, CM - color moments and VW - bag-of-visual-words.

	AV	PV	IV	CUV	FAV
CH	0.068	0.052	0.017	0.042	0.029
CM	0.096	0.039	0.013	0.031	0.032
VW	0.105	0.068	0.046	0.071	0.117
CH+CM	0.111	0.076	0.034	0.046	0.034
CH+CM+VW	0.242	0.131	0.087	0.165	0.121

search is more challenging than general content-based image retrieval tasks. The mechanism of content-based retrieval method decides it can only return visually similar images (w.r.t the query image) in machine’s view. While in landmark image retrieval, the positive results should represent or contain *visual appearance of the targeted landmark*, which implies that a result could be negative even it has similar visual content with the query image.

Queries with the whole view of a landmark are expected to get better results than partial views (*PV*) and interior views (*IV*). Queries of three types (views of different angles (*AV*), close-up views (*CUV*) and far-away views (*FV*) contain the whole exterior view of a landmark. In general, *CUV* contains more details and *FV* contains foreground and background objects (e.g. pedestrians and vehicles). These extra details and objects increase the difficulty of retrieval based on visual features. As a result, *AV* obtains the best results among them. Surprisingly, *PV* gets similar performance as *CUV*. We find that it is because that the captured partial views are usually special scenes or objects which tend to attract more attentions, resulting in more images about them. The search performance of *IV* is the worst, which is in our expectation. Because interior views typically contain more objects and complicate structures, and the lighting conditions vary greatly at different time or from different angles. Comparing the performance of different landmarks, we found that the queries of *Marina Bay* and *Merlion Park* obtain much better results than queries of other landmarks, which are all buildings. It implies that buildings are more difficult to search based on visual features. And it can be explained by the viewpoint in [3]: “*buildings tend to have few discriminative visual features and many repetitive structures*”.

Efficiency Study. As a baseline study, we have not used any indexing method in the experiments. For single visual feature, the computation time is spent on computing and sorting the similarity scores of all images in the subset; for the combination of different features, there are two additional time-consuming steps - the normalization and summation of computed scores of individual features. Obviously, higher dimensionality of visual features needs longer processing time. The results are shown in Fig. 4. From the results, we can see that the slight improvement on search performance by combining visual features pays high time cost. The 1289-dimensional combined features take 4.14s per query, which is much longer than the computation time of 1000-dimensional visual words (1.19s).

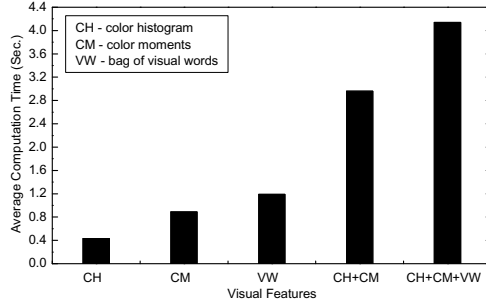


Fig. 4. Computation time per query using different visual features

Table 5. Hierarchy structure and details of selected landmarks in Singapore (for the types, please refer to section 2.4)

City	Districts	Types	Landmarks	Image Number
Singapore	Marina Area	B3	Boat Quay	2887
			Marina Bay	1108
		A6	Merlion Park	3070
	CBD	B6	Armenian Church	3232
	City Hall	B1	City Hall	2805
		B5	Raffles City	3401
			National Museum	3128
	Beach Road	B6	Cathedral of the good Shepherd	677
			Church of Our Lady of Lourdes	2887
			Church of Saints Peter and Paul	3279
	Tanglin	A6	Singapore Botanic Gardens	3116
	Newton	A4	Bukit Timah Nature Reserve	2861
	Orchard	A6	Istana Park	728
		B7	Orchard Road	2888
	Changi	B7	Changi Airport	2762
		A2	Pulau Ubin	2363
	Far North	A5	Night Safari Singapore	2834
			Singapore Zoo	2755
	Jurong	A5	Jurong Bird Park	3333
		A4	Sungei Buloh Wetland Reserve	2946
	Hougang	B6	Church of the Nativity of the Blessed Virgin Mary	589
	Keppel	A5	Butterfly Park & Insect Kingdom	2088
			Harbourfront center	2863
Kusu Island			2986	
Mount Faber			2766	
Santosa			2137	
	B1	Universa Studios Singapore	2740	
South West	A6	Haw Par Villa	2510	

5 Conclusion and Future Work

In this paper, we introduce the construction of a large scale mobile landmark image dataset. The dataset contains various kinds of textual features and includes six types of visual features. Based on the dataset, we identified and discussed several closely related research issues. We also conducted a set of experimental studies on mobile landmark search using different visual features. The search

accuracy of different visual queries and the search efficiency of different visual features were reported and comprehensively analyzed. The results show the weakness of content-based landmark search. Both search accuracy and efficiency need to be improved. These outcomes can be used as baselines to facilitate the future related research. Further, the dataset can be applied as the testbed to assess performance of different mobile search platforms.

References

1. <https://sites.google.com/site/smuzycheng/landmark-dataset>
2. Avrithis, Y., Kalantidis, Y., Tolias, G., Spyrou, E.: Retrieving landmark and non-landmark images from community photo collections. In: ACM MM (2010)
3. Chen, D.M.: City-scale landmark identification on mobile devices. In: CVPR (2011)
4. Chen, W.-C., Battestini, A., Gelfand, N., Setlur, V.: Visual summaries of popular landmarks from community photo collections. In: ACM MM (2009)
5. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.* 40(2) (2008)
6. Gao, Y., Wang, M., Zha, Z., Shen, J., Li, X., Wu, X.: Visual-textual joint relevance learning for tag-based social image search. *IEEE Transactions on Image Processing* (inprint)
7. Huang, J., Kumar, S., Mitra, M., Zhu, W.-J., Zabih, R.: Image indexing using color correlogram. In: CVPR (1997)
8. Ji, R., Duan, L.-Y., Chen, J., Yao, H., Yuan, J., Rui, Y., Gao, W.: Location discriminative vocabulary coding for mobile landmark search. *International Journal of Computer Vision* 96(3) (2012)
9. Kennedy, L., Naaman, M.: Generating diverse and representative image search results for landmarks. In: WWW (2008)
10. Li, Y.P., Crandall, D.J., Huttenlocher, D.P.: Landmark classification in large-scale image collections. In: ICCV (2009)
11. Lowe, D.: Distinctive image features from scale-invariant keypoints. *Computer Vision* 2(60), 91–110 (2004)
12. Manjunath, B.S., Ma, W.Y.: Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell.* 18(8), 837–842 (1996)
13. Park, D.K., Jeon, Y.S., Won, C.S.: Efficient use of local edge histogram descriptor. In: ACM MM (2000)
14. Ren, Y.H., Yu, M., Wang, X.J., Zhang, L., Ma, W.Y.: Diversifying landmark image search results by learning interested views from community photos. In: WWW (2010)
15. Shapiro, L.G., Stockman, G.C.: *Computer Vision*. Prentic Hall (2003)
16. Simon, I., Snavely, N., Seitz, S.M.: Scene summarization for online image collections. In: ICCV (2007)
17. Wang, M., Ni, B., Hua, X.-S., Chua, T.-S.: Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Comput. Surv.* 44(4), 25 (2012)
18. Wang, M., Yang, K., Hua, X.-S., Zhang, H.: Towards a relevant and diverse search of social images. *IEEE Transactions on Multimedia* 12(8), 829–842 (2010)

Geographical Retagging

Liujuan Cao¹, Yue Gao², Qiong Liu³, and Rongrong Ji⁴

¹ Harbin Engineering University, Harbin, 150001, P.R. China

² National University of Singapore, 117417, Singapore

³ Huazhong University of Science and Technology, Wuhan, 430074, China

⁴ Columbia University, NY State, 10027, United States

{caoliujuan, kevin.gao}@gmail.com,

rrji@ee.columbia.edu

Abstract. While the geographical tag has brought a novel insight into the multimedia content analysis and understanding, how to improve the tagging accuracy has been rarely exploited. In this paper, we present a novel geographical retagging algorithm to improve the inaccurate geographical tags from an automatic photo content based association and refinement perspective. We do not resort to the time-consuming camera pose estimation and scene geometry recovery schemes like structure-from-motion. Instead, our algorithm is deployed based on a very simple neighbor statistical significance test, i.e., geographically nearby images, if near duplicate, should follow a more smooth affine transform comparing with those farther away. Such an assumption is robust to noisy photo contents caused by multiple factors, such as indoor/outdoor changes, occlusions, or viewing angle changes. It is also very fast comparing to alternative approaches like structure-from-motion or simultaneous localization and matching. We have shown the accuracy, efficiency, and robustness of the proposed retagging algorithm for refining the geographical tags of Flickr images.

Keywords: Social Media, Geographical Tagging, Tag Refinement, Statistical Significance Testing, Flickr Images.

1 Introduction

Nowadays there is an explosive growth of geographical tags in social multimedia. For instance, photo sharing websites such as Flickr and Picasa enables users to tag the geographical locations, i.e. their latitudes and longitudes, of their uploaded photos on the map. Together with the metadata context and the visual contents, such massive geographical tags provide a novel viewpoint to organize, browse, and summarize the community contributed photo collections to “see” the world. One example is to mine representative landmarks for the purpose of touristic recommendation, as recently investigated in [1][4].

The geographical tags of social multimedia come from multidiscipline. For instance, nowadays many mobile devices are equipped with digital cameras and GPS, which enables the mobile users to tag the geographical location right after

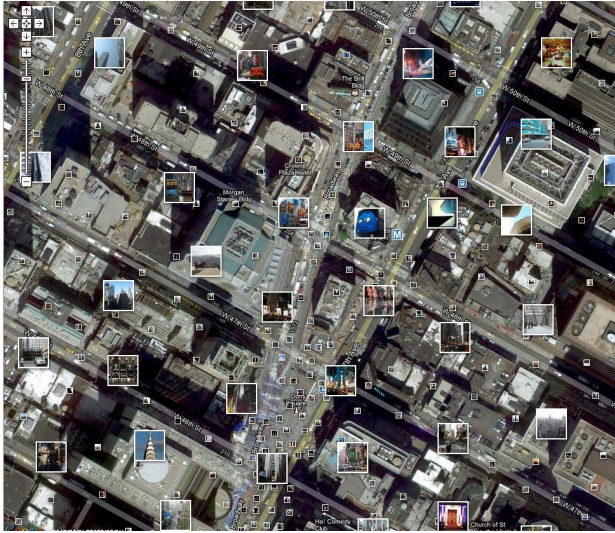


Fig. 1. An example of the inaccurate geographical tagging in Time Square, where some of the geographically tagged photos are with more than 1-block distortion

photo capturing. Another source comes from the geographical tagging interfaces provided in photo sharing websites such as Flickr and Pananomia, where users can tag pins of their photos on the geographical map. In the automatic circumstance, locations of a given webpage can be inferred through the textual location descriptions within this webpage, such as city names and zip codes, typically achieved by using a hierarchical matching over a location gazetteer. Visual based matching is another alternative towards automatic geographical tagging, for example some large-scale global image locating engines developed based on near-duplicate visual search [4][22].

The geographical tag plays an important role in knowledge mining and instance recognition from social multimedia. For instance, a commonsense component in preprocessing geo-tagged multimedia is to leverage the geographical tags for scene/landmark partition. In some other circumstances such as landmark search or location recognition, the geographical tag associated with the reference images are treated as the ground truth to be returned by location search engines [20][22][5]. For applications such as city-scale scene reconstruction, the ability to precisely locate the reference images in the geographical map not only ensures the reconstruction accuracy, but also improves the reconstruction efficiency.

Without regard to its fundamental importance, the geographical tag is not always accurate. For instance, the GPS functionality in the mobile device is occasionally distorted in the urban areas. And the user tagging in Flickr or Pananomia is also imprecise, due to both the map resolution and the user mis-operation. Figure 1 shows an example of the inaccurate geographical tagging on the landmark of Time Square, where most of the photo are incorrectly located comparing to their actual geographical locations. However, a systematic study of the accuracy distortion of

the geographical tagging is left unexploited. Nevertheless to say, manual rectification is too time-consuming and therefore cannot be scaled up for coping with the dramatic data growth. Therefore, efforts toward an automatic and unsupervised refinement of these geographical tags can largely benefit both the academy and the industry. However, it is left unexploited in the existing works.

In this paper, we propose an approach to automatically discover and refine imprecise geographical tags in a scalable manner. Our principle is to make use of the visual content correspondence among these images as an evidence, from which rearrange their location to seek an optimal visual consistency. To this end, we argue that the existing works in automatic image calibration as well as scene geometry reconstruction too computation intensive to be scaled up. For instance, while using structure-from-motion can automatically calibrate the image locations as well as camera poses, their bundle adjustment based feature correspondence is very time-consuming, therefore “over-qualified” and inefficient. Instead, we propose an efficient yet effective geographical retagging algorithm based upon a neighbor statistical significance test, e.g., geographically nearby images, if near-duplicate, should follow a smooth spatial transform among their visual descriptor matching. Such an assumption is also not too constraint comparing to scene reconstruction techniques, while still being robust to visually non-overlapped image caused by occlusions or viewing angle changes. In implementation, we can directly adopt a spatial sliding window over the geographical map to largely reduce the computational complexity. Experimental results show that, our scheme can achieve a large speedup while retaining comparable tagging precision comparing to time-consuming scene reconstruction techniques.

Section 2 reviews related work in using and refining the geographical tags. Section 3 gives a close look at the geographical tagging precision. Section 4 gives detailed explanation of our approach. Experimental results are given in Section 5. Finally, we conclude in Section 6 and point out our future work.

2 Related Work

Canonical View Selection. Aiming at unsupervised landmark mining, Kennedy and Naaman [1] presented a tag-location-vision processing flowchart to group views from geo-tagged Flickr photos of San Francisco. Metadata consensus is considered by weighting author confidences, time durations, and context variations. However, the community character is not thoughtfully considered in [1]. And the photo representativity is only measured by counting local feature associations in each visual cluster. Recently, Ji et al. [3] presented a graph-based mining framework to discover landmarks from blogs. Cao et al. [2] proposed to model the canonical correlation between heterogeneous features and an annotation lexicon of interest, and built a generalized semantic and geographical annotation engine.

Structure from Motion. Recently, Irschara et al. [5] adopted structure-from-motion technique to build 3D scene models for vision-based city scene localization, which has been combined with vocabulary tree-based visual indexing for simultaneously city scene modeling and real-time location recognition. Related

works also exist in robotic vision research, in which the visual SLAM are usually adopted for indoor robot self-localization in small-scale environments [6], which was also used in multi-view object search [17][18][19].

Landmark Search. Towards city-scale landmark search and recognition, Schindler et al. [20] presented a location recognition system through geo-tagged video streams with multiple path search in the vocabulary tree. Our previous works in [21] proposed a density-based metric learning to optimize the hierarchical structure of vocabulary tree for street view location recognition. Cristani et al. [22] learnt a global-to-local image matching for location recognition. And their consecutive work in [23] identified landmark buildings based on image data, metadata [10][11][12], and other photos taken within a consecutive 15-minute window. In addition, Irschara et al. [5] further leverage structure-from-motion (SFM) to build 3D scene models for street views, combined with vocabulary tree for simultaneously scene modeling and location recognition. Xiao et al. [6] proposed to combine bag-of-features with simultaneous localization and mapping (SLAM) to further improve the recognition precision. The quantization issues in visual vocabulary are recently also well addressed to fit the city-scale landmark search scenario, such as the works in [21][26]. Incrementally vocabulary indexing is also explored in [9] to maintain a landmark search system in a time varying database, which benefits also other search techniques such as [8][14][13][15].

Towards worldwide landmark search and recognition, the IM2GPS system [24] inferred possible location distributions of a given query by visual matching in a worldwide, geo-tagged landmark dataset. As a consecutive work, Kalogerakis et al. further [25] demonstrated how to combine single image matching with sequential data to improve matching accuracy. Zheng et al. [4] developed a worldwide landmark recognition system, which used a predefined landmark list to query online image search engines to selected candidate images, followed by re-clustering and pruning to locate the final landmark location. Recent works also proposed to mine representative landmarks worldwide, such as using sparse representation [16][27].

3 A Close Look at Geographical Tagging

Nowadays, many mobile devices, such as cell phones, Personal Digital Assistant (PDA), and digital cameras, are integrated with the GPS (Global Position System) modules. Such integration makes it possible to precisely tag GPS information (latitude and longitude) on the multimedia metadata. In addition, in the case that both multimedia content and GPS trajectory are simultaneously recorded with time stamps, GPS tagging can be also bound to multimedia content by associating their time stamps [7][3].

Explicit GPS tagging also comes from photo sharing websites such as Flickr and Picasa. Many of such websites enable users to tag (precisely or approximately) the geographical locations of their uploaded photos on the map. For instance, the geographical tagging interface was launched by Flickr in 2006, and Flickr has reported over 90 million geo-tagged photos in January 2009. The photos with GPS information are increasing even faster these days.

Table 1. A case study of geographical tagging precision in Flickr

Distortion	<10m	10m-20m	20m-30m	30m-40m	40m-50m	50m-100m	>100m
Percentage (%)	1.25	2.45	2.204	4.836	3.830	2.736	6.072

However, the correctness of these tagging is not always guaranteed. For example, Table 1 gives an overview of the geographical tagging precision from the Flickr geo-tagged images. To accomplish this case study, we asked a group of volunteers to label the actual position for each photo. Then, we adopt the mean latitude and longitude to represent the ground truth location for this image¹. Then, the difference between this “ground truth” geo-tag and the Flickr geo-tag is compared. It is obvious that such tagging is actually not as precise as we expected: The Flickr’s geographical tagging comes from both interactive Flickr user labeling through their web interface, or the mobile uploading (from the EXIF file) detected by the GPS sensor. Both are actually very imprecise.

4 Geographical Retagging

4.1 Problem Formulation

Given a sliding window working on the geographical map, we denote the collection of geographically tagged images located in this sliding window as $\mathbf{I} = \{I_1, I_2, \dots, I_n\}$, where each image I_i is tagged with its geographical coordinate $\langle La_i, Lo_i \rangle$. Our goal is to refine $\langle La_i, Lo_i \rangle$ into $\langle La'_i, Lo'_i \rangle$ for each I_i , expecting that a more precise latitude and longitude coordinate pair can be derived after retagging.

4.2 Statistical Significance Test

We resort to the statistical significance test, or so-called “statistical hypothesis test”, to achieve this goal. A statistical hypothesis is an assumption about a population parameter². This assumption may or may not be true. As a result, the hypothesis test is defined as the procedures used by statisticians to accept or reject any given statistical hypothesis.

The best way to determine whether a statistical hypothesis is true is to examine the entire population. However, this is often impractical due to the scalability issue, therefore an alternative approach is to examine a random sample set from the population. If the sample data are not consistent with the statistical hypothesis, the hypothesis would be rejected and subsequently a new hypothesis would be generated. We define the initial and regenerated hypothesis as:

- **Null hypothesis.** The null hypothesis, denoted by H_0 , is defined as the hypothesis that samples observations purely from chance.

¹ Once the volunteer is not sure about the actual geographical location of this image, he or she can label as “unknown”.

² <http://stattrek.com/hypothesis-test/hypothesis-testing.aspx>

- **Alternative hypothesis.** The alternative hypothesis, denoted by H_1 , is defined as the hypothesis that samples observations are influenced by some non-random cause.

Given a decision function with choices C_a and C_b and their priors $P(C_a)$ and $P(C_b)$ respectively, a null hypothesis might be an observation that follows their distribution priors, i.e., whether $\frac{p(C_a|I_i)}{p(C_b|I_i)}$ is identical to $\frac{p(C_a)}{p(C_b)}$, where I_i is the current observation that corresponds to a given geographically tagged photo. On the other hand, the alternative hypothesis might be an observation that with a very different distribution of $\frac{p(C_a|I_i)}{p(C_b|I_i)}$ comparing to $\frac{p(C_a)}{p(C_b)}$. Suppose we have n observations corresponding to $\mathbf{I} = \{I_1, I_2, \dots, I_n\}$, and we have:

$$P_{Observation} = \frac{p(C_a|\mathbf{I})}{p(C_b|\mathbf{I})} = \frac{\sum_{i=1}^n p(C_a|I_i)}{\sum_{i=1}^n p(C_b|I_i)} \quad (1)$$

Given $P_{Observation}$, if it conflicts with the prior distribution $\frac{p(C_a)}{p(C_b)}$, we would be inclined to reject the null hypothesis H_0 . In such a case, we will need to re-draw the alternative hypothesis H_1 as H_0 and re-test again, which is iterated until: (1) the iteration is over a given number; or (2) the current hypothesis is accepted. The following pipeline outlines the procedure of the proposed hypothesis testing:

- State the hypothesis, which involves stating the null and alternative hypothesis. The stated hypothesis should be mutually exclusive, i.e., if one is true, the other must be false.
- Formulate an analysis plan, which describes how to use sample data to evaluate the null hypothesis. The evaluation often focuses around a single test statistic.
- Analyze sample data, which finds the value of the test statistic, for instance the *mean score*, *proportion*, *t-score*, *z-score*, etc., as described in the analysis plan.
- Interpret results, which applies the decision rule described in the analysis plan: If the value of the test statistic is unlikely, based on the null hypothesis, reject and replace the null hypothesis by the alternative hypothesis.

4.3 Geographical Retagging by Hypothesis Testing

Given the above hypothesis testing algorithm, there are three components to be determined to bridge such testing into our geographical retagging scenario: (1) $P_{Observation}$ definition, which refers to defining the correlation between image pairs within \mathbf{I} . (2). Hypotheses statement, which would be formulated as selecting one under-evaluated photo’s geographical tag; (3). Analysis plan, which would be correspond to an optimal decision order to fit $\{I_1, I_2, \dots, I_n\}$ into the relative locations of rest images based on their location transform based on the hypothesis defined in Step (1); (4). Switching between null and alternative hypothesis, such that the efficiency is ensured. We detail each components as follows:

Algorithm 1. The Statistical Significance Testing Procedure for Geographical Retagging

- 1 **Input:** Image collection $\mathbf{I} = \{I_1, I_2, \dots, I_n\}$ and their geographical tags $\{\langle La_i, Lo_i \rangle\}_{i=1}^n$.
 - 2 **Output:** The refined geographical tags $\{\langle La'_i, Lo'_i \rangle\}_{i=1}^n$.
 - 3 **Pairwise Transform:** Calculate the pairwise transform of any I_i and I_j pair, forming the connection graph $\{\mathbf{G}_{i,j}\}_{i,j \in N}$.
 - 4 **for** $I_i \in \mathbf{I}$ **do**
 - 5 | **Statistical Significance Testing** under the hyper thesis $I_i = I_{H_0}$.
 - 6 | Update the Testing threshold T if this is a more significant hypothesis.
 - 7 **end**
 - 8 Refine $\{\langle La_i, Lo_i \rangle\}_{i=1}^n$ based on the survived statistical hypothesis.
-

P_{Observation} Definition. To define $P_{Observation}$, we introduce the following perspective transform matrix as the basic connection to bridge each image pair I_i and I_j within $\{I_1, I_2, \dots, I_n\}$:

$$\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = \begin{pmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{pmatrix} \begin{pmatrix} x_j \\ y_j \\ z_j \end{pmatrix} \quad (2)$$

Therefore, any two images are connected based on their respective transform matrix, allowing as $XYZ(I_i) = \mathbf{M}_{i,j}XYZ(I_j)$, where $XYZ(I_i)$ is the physical position in the world coordinate and is simplified as $\langle La, Longitude \rangle$ without specifying any altitudes. Now, we can image a connection graph $\{\mathbf{G}_{i,j}\}_{i,j \in n}$, where each entry at (i, j) corresponds to their transform $\mathbf{M}_{i,j}$, which means $G_{i,j} = G_{i,j}^t$.

Hypotheses Statement. For each i -th image, we have a geographical tagging pair $\langle La_i, Lo_i \rangle$, which is defined as the hypothesis candidate. To initialize our selection, we choose the following image from \mathbf{I} that satisfies:

$$I_{H_0} = \arg \max_i \sum_j Confidence(\mathbf{M}_{i,j}) \quad (3)$$

which selects the image geographical location with the maximal confidence following the transform, where $Confidence(\mathbf{M}_{i,j})$ denotes the confidence of transforming, which is calculated as matching pair consistency scores such as using RANSAC.

Analysis Plan. Subsequently, we test the rest images within $\{I_1, I_2, \dots, I_n\}$ except I_{H_0} to analyze the hypothesis correctness of H_0 . To that effect, we testify for each I_i and the hypothesis image I_{H_0} the following transform:

$$\langle La'_i, Lo'_i \rangle = G_{i,H_0} \langle La_{H_0}, Lo_{H_0} \rangle \quad (4)$$

Then the overall significance test score for I_{H_0} is defined as:

$$P(I_{H_0} | \{I_1, I_2, \dots, I_n\} - I_{H_0}) = \sum_{i \in [1, N], i \neq H_0} \|\langle La'_i, Lo'_i \rangle - \langle La_i, Lo_i \rangle\|_2 \quad (5)$$

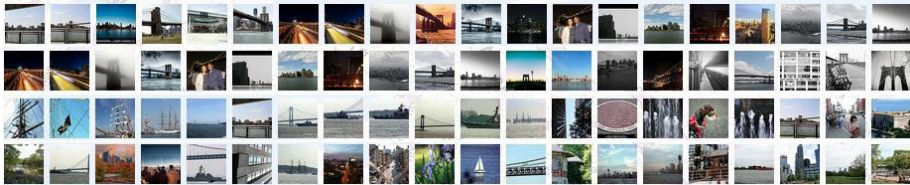


Fig. 2. A snapshot of geo-tagged Flickr photos near Brooklyn Bridge in New York City

where subsequently a threshold T , defined as the currently best hypothesis, is used to reject or accept a hypothesis H_0 .

Finally, given the accepted H_0 , I_i 's geographical tag with $Confidence(\mathbf{M}_{i,H_0})$ less than all the rest images is defined as the true geographical tag. Subsequently, all other geographical tags in $\{I_1, I_2, \dots, I_n\} - I_{H_0}$ is adapted using Equation 4 to come up with their retagging results. This procedure ensures that we do not force any visually unreliable matching, which is caused by indoor-outdoor images, occlusion, partial matching, and mis-tagging from other facet of the same landmark, or simply from different landmarks.

Switching between Null and Alternative Hypotheses. To ensure an efficient switching between the null and alternative hypothesis, we use an empirical rule: For a given hypothesis, given $P(I_{H_0} | \{I_1, I_2, \dots, I_n\} - I_{H_0})$, once it is reject, we rank the violation scores $\| \langle La'_i, Lo'_i \rangle - \langle La_i, Lo_i \rangle \|_2$ for each I_i . Then we select the medium I_j as the new hypothesis H_1 . Algorithm 1 outlines the overall procedure for our hypothesis testing.

5 Experimental Validations

To evaluate our proposed geographical retagging algorithm, we deploy our algorithm on an application to refine the geo-tags of Flickr images in this section. We report our improvements in both the tagging accuracy and the run-time efficiency, with quantitative comparisons to a group of alternative approaches.

5.1 Data Collection and Ground Truth Labeling

We collect over 10 million geo-tagged photos from photo sharing websites of Flickr and Panoramio. We crawled photos from five worldwide metropolitans including *Beijing*, *New York City*, *Barcelona*, *Singapore* and *Florence*.

Since it is infeasible to manually label all 10 million images as we did in our case study (Section 3), we manually select 10 landmarks in each city for labeling, which results in 50 groups of landmark image collections. We then ask a group of volunteers to label the ground truth geographical location of each image. Similar to our case study, for each reference image, we ask each volunteer to label the best geographical location from his or her perspective. Then, we adopt the mean latitude and mean longitude to represent the ground truth location for this image. The difference between the manually labeled geographical tag and the geographical tag provided by the Flickr API is compared.

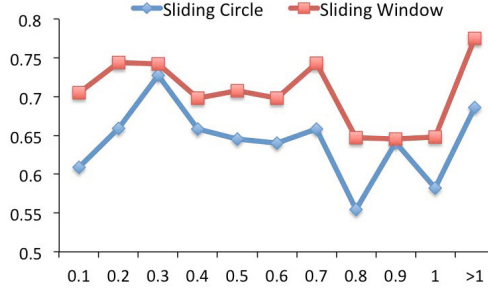


Fig. 3. The geographical retagging precision with respect to different radiuses of the geographical sliding window

5.2 Evaluation Protocols

We run our geographical retagging algorithm over these 10 million geo-tagged photos, using a sliding window over the geographical map for each city. The window radius is fixed to be D , which is tuned subsequently in Figure 3. We then measure the geographical retagging accuracy using average distance error of individual images, and the run-time speed with a regular PC.

5.3 Quantitative Results

Figure 3 shows the distance error of all these 50 landmark regions with respect to the geographical sliding window radius D . It is obvious that such radius does not impose a significant effect to the average distance error of individual images, once it is over 0.36 kilometer and is less than 1.24 kilometer. For the consideration of sliding window scanning efficiency, the latter of which is subsequent leveraged as our final sliding window radius.

Figure 4 shows the effect of choosing different initial hypothesis based on: (1) the proposed minimal violation rule as in Equation 3, (2) randomized initialization, and (3) sequential initialization, all of which are with respect to the labeled 10 landmarks in New York, including *Brooklyn Bridge*, *Central Synagogue*, *Chrysler Building*, *Eldridge Street Synagogue*, *Empire State Building*, *Flatiron Building*, *New York Public Library*, *New York Stock Exchange*, *Plaza Hotel*, *St. Patrick’s Cathedral*. As can be seen, our final approach (1) slightly outperformed the straightforward alternatives as in (2) and (3) in the visually more consistent landmarks such as *Flatiron Building* and *Plaza Hotel*. However, for those “hard” landmarks such as *Brooklyn Bridge* and *Eldridge Street Synagogue*, our solution (1) performs much better than its alternatives (2) and (3).

Table 2 shows the processing time of our geographical retagging before and after our heuristic ordering. In the former case, we directly change the H_0 based on the orders of $\{I_1, I_2, \dots, I_n\}$, which, as shown in Table 2, involves long processing time overload, compared with our scheduling.

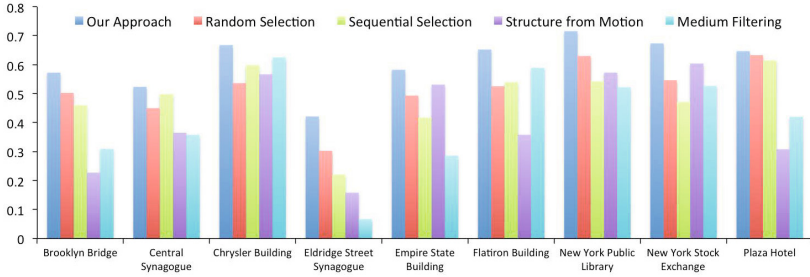


Fig. 4. Case study of geographical retagging precision

Table 2. Time cost with respect to different components

Approach	Components	Pairwise Matching	Stat. Sig. Test	Retagging
Selective	(Second)	$20.78 \times n$	6.25	3.80
Random	(Second)	$20.78 \times n$	17.58	3.63
Sequential	(Second)	$20.78 \times n$	15.64	4.15

6 Conclusion and Future Work

With the proliferation of social multimedia, its geographical tag accuracy imposes an increasing importance towards its precise mining and recognition. However, how to refine the initial imprecise geographical tags is left unexploited. In this paper, different from resorting to the time-consuming structure-from-motion like image position adjustment, we introduce a novel statistical significance testing approach towards an efficient yet robust geographical retagging, i.e., geographically nearby images, if near duplicate, should follow a more smooth affine transform comparing with those farther away. Our approach has shown extremely excellent performance in 10 million Flickr images. In our future work, we will focus on replacing the estimation of transform matrix between image pairs to further improve the geographical retagging accuracy.

Acknowledgement. This work was supported by National Natural Science Foundation of China (NSFC) (No.61170194 and 61202301).

References

1. Kennedy, L., Naaman, M., Ahern, S.: How flickr helps us make sense of the world: context and content in community contributed media collections. *ACM Multimedia (2007)* 1,3
2. Cao, L.-L., Yu, J., Luo, J., Huang, T.S.: Enhancing Semantic and Geographic Annotation of Web Images via Logistic canonical correlation regression. *ACM Multimedia (2009)* 3

3. Ji, R., Xie, X., Yao, H., Ma, W.-Y.: Mining city landmarks from blogs by graph modeling. *ACM Multimedia* (2009) 3, 4
4. Zheng, Y.-T., Zhao, M., Song, Y., Adam, H.: Tour the world: building a web-scale landmark recognition engine. In: *CVPR* (2009) 1, 2, 4
5. Irschara, A., Zach, C., Frahm, J., Bischof, H.: From structure-from-motion point clouds to fast location recognition. In: *CVPR* (2009) 2, 3, 4
6. Xiao, J., Chen, J., Yeung, D.-Y., Quan, L.: Structuring Visual Words in 3D for Arbitrary-View Object Localization. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part III*. LNCS, vol. 5304, pp. 725–737. Springer, Heidelberg (2008) 4
7. Jia, M., Fan, X., Xie, X., Li, M., Ma, W.-Y.: Photo-to-search: Using camera phones to inquire of the surrounding world. In: *MDM* (2006) 4
8. Ji, R., Yao, H., Wang, J., Xu, P., Liu, X.: Clustering-based subspace SVM ensemble for relevance feedback learning. In: *ICME* (2008) 4
9. Ji, R., Xie, X., Yao, H., Wu, Y., Ma, W.-Y.: Vocabulary tree incremental indexing for scalable location recognition. In: *ICME* (2008) 4
10. Wang, M., Ni, B., Hua, X.-S., Chua, T.-S.: Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Computing Surveys* (2012) 4
11. Wang, M., Hong, R., Li, G., Zha, Z.-J., Yan, S., Chua, T.-S.: Event driven Web Video Summarization by Tag Localization and Key-Shot Identification. *TIP* (2012) 4
12. Wang, M., Hong, R., Yuan, X.-T., Yan, S., Chua, T.-S.: Movie2Comics: Towards a Lively Video Content Presentation. *TMM* (2012) 4
13. Ji, R., Duan, L.-Y., Chen, J., Yao, H., Yuan, J., Rui, Y., Gao, W.: Location Discriminative Vocabulary Coding for Mobile Landmark Search. *IJCV* (2012) 4
14. Ji, R., Yao, H., Liu, W., Sun, X., Tian, Q.: Task Dependent Visual Codebook Compression. *TIP* (2012) 4
15. Ji, R., Duan, L.-Y., Yao, H., Xie, L., Rui, Y., Gao, W.: Learning to Distribute Vocabulary Indexing for Scalable Visual Search. *TMM* (2012) 4
16. Ji, R., Gao, Y., Zhong, B., Yao, H., Tian, Q.: Mining City Landmarks by Modeling Reconstruction Sparsity. *TOMCCAP* (2011) 4
17. Gao, Y., Tang, J., Hong, R., Dai, Q., Chua, T., Jain, R.: W2Go: A Travel Guidance System by Automatic Landmark Ranking. *ACM Multimedia*, 123–132 (2010) 4
18. Gao, Y., Wang, M., Tao, D., Ji, R., Dai, Q.: 3D Object Retrieval and Recognition with Hypergraph Analysis. *TIP* (2012) 4
19. Gao, Y., Wang, M., Zha, Z., Tian, Q., Dai, Q., Zhang, N.: Less is More: Efficient 3D Object Retrieval with Query View Selection. *TMM* (2011) 4
20. Schindler, G., Brown, M.: City-scale location recognition. In: *CVPR* (2007) 2, 4
21. Ji, R., Xie, X., Yao, H., Ma, W.-Y.: Mining city landmarks from blogs by graph modeling. *ACM Multimedia*, 105–114 (2009) 4
22. Cristani, M., Perina, A., Castellani, U., Murino, V.: Geolocated image analysis using latent representations. In: *CVPR* (2008) 2, 4
23. Crandall, D., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world's photos. In: *WWW* (2009) 4
24. Hays, J., Efros, A.: IMG2GPS: estimating geographic information from a single image. In: *CVPR* (2008) 4
25. Kalogerakis, E., Vesselova, O., Hays, J., Efros, A., Hertzmann, A.: Image sequence geolocation with human travel priors. In: *CVPR* (2009) 4
26. Ji, R., Yao, H., Xie, X., Tian, Q.: Vocabulary Hierarchy Optimization and Transfer for Scalable Image Search. *IEEE MM* (2011) 4
27. Ji, R., Xie, X., Yao, H., Ma, W.-Y.: Mining City Landmarks by Graph Modeling. *ACM Multimedia* (2009) 4

Recompilation of Broadcast Videos Based on Real-World Scenarios

Ichiro Ide

Graduate School of Information Science, Nagoya University,
1 Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan
ide@is.nagoya-u.ac.jp

Abstract. In order to effectively make use of videos stored in a broadcast video archive, we have been working on their recompilation. In order to realize this, we take an approach that considers the videos in the archive as video materials, and recompiling them by considering various kinds of social media information as “scenarios”. In this paper, we will introduce our works in news, sports, and cooking domains, that makes use of Wikipedia articles, demoscopic polls, twitter tweets, and cooking recipes in order to recompile video clips from corresponding TV shows.

1 Introduction

In recent years, following the increase in the capacity of storage devices, research on retrieval and browsing of videos stored in a large-scale broadcast video archive has become active. When considering the retrieval of video media compared to text and image media, there is a problem that it is more difficult to browse and conceive the retrieved information at a glance.

Considering this problem, we have been working on methods that do not simply provide to the users, a list of individual video clips in the archive as the retrieved results, but instead, provide a presentation of recompiled video clips according to “scenarios” obtained from the real world; various kinds of social information available on, mostly, but not limited to, the Web.

The rest of the paper is organized in three parts, where we will introduce our works based on the above approach in the domains of news, sports, and cooking.

2 Applications to News Contents

For news contents, we will introduce two different works on video recompilation that makes use of different kinds of social information as “scenarios”.

2.1 Description of Wikipedia Articles with Videos

- Video: News show
- Scenario: Wikipedia articles on current topics

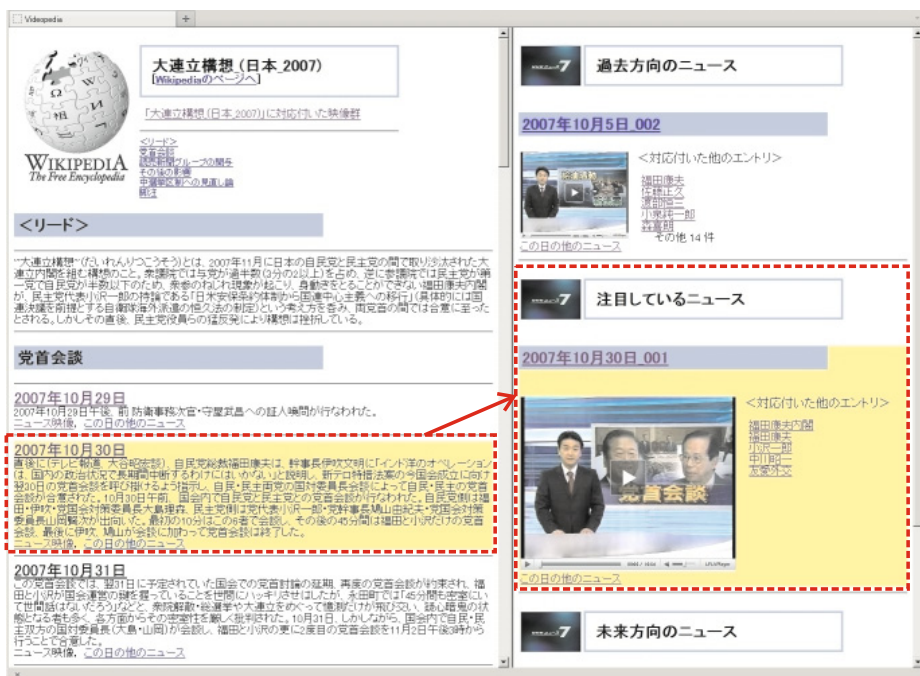


Fig. 1. The “Videopedia” interface. The left side of the screen shows the original chronological explanation texts in Wikipedia, while the right side shows the video clip corresponding to the explanation text high-lighted in the left-side. Video clips of news stories preceding and succeeding the corresponding news story along the topic thread structure are also shown above and below. In addition, links to other corresponding Wikipedia articles are listed next to each video clip.

Wikipedia is an online encyclopedia that allows the general public to edit at any time, so it usually contains up-to-date information on various phenomena in the real world. Here, we considered each Wikipedia article as a dynamic scenario, and proposed a method that visually describes it with the help of a sequence of video clips from the news video archive.

In order to realize this, we proposed a framework that links video clips from news shows (news stories) to chronological explanations texts in Wikipedia articles on current topics, and developed an interface “Videopedia” (Fig. 1) to demonstrate the results. Details of the method could be found in [1].

Linking Video Clips (News Stories) to Wikipedia Articles. First, we need to link corresponding video clips to Wikipedia articles. Here, each video clip represents a news story.

In order to narrow-down the candidates, date expressions are extracted from the chronological explanation in a Wikipedia article, and only video clips broadcast on or around the date are analyzed.

Then, the most related video clip is selected by comparing the term frequencies in the Wikipedia article and the closed-caption of the selected video clips. This process is performed for each date expression extracted.

Interpolation Based on the Topic Thread Structure. Since the detailed-ness of the descriptions are different in news shows and Wikipedia articles, the above process is not sufficient to obtain links to most of the date expressions. In order to compensate for this problem, the topic thread structure proposed in [2] was used to interpolate the links obtained in the above process.

The precision and the recall of the links obtained from the above processes were 82.1% and 75.8%, respectively, in a manual evaluation of three Wikipedia articles.

2.2 Generation of a Summarized Video on a Specific Person in News

- Video: News show
- Scenario: Demoscopic polls

In broadcasting stations, there are often cases that they need to produce a video that introduces a person’s personal history. In most cases, such a necessity arises suddenly before the news show, so the producers need to gather source material and compile them in a limited period of time.

To provide an automatic solution for such a task, we focused on a “Prime Minister” as an example, since he appears in TV news quite frequently, and proposed a method to provide a biased summarized video that explains why he had to resign in the end. In order to produce such an explanation, we collected video clips corresponding to major events that occurred while he was in office, by referring to demographic polls and also features obtained from topic thread structures. In order to detect major events, we prepared two approaches: Template-based, which detects typical events for all Prime Ministers, and Topic-based, which detects major events specific to the period. Details of the method could be found in [3].

Preparation of Source Video Clips (News Stories). First, as the initial dataset, news stories that contain the Prime Minister’s name as a subject in the closed-caption during his period in office are extracted.

Template-Based News Story Selection. In the inauguration / resignation periods (i.e. the beginning and the end of his period in office), there are typical events such as inauguration / resignation speeches, visits to foreign countries to see foreign leaders, and so on. In order to detect such stories, we prepared templates composed of typical keywords, and searched for them in the closed-caption of the stories broadcasted in the beginning and the end of his period in office.

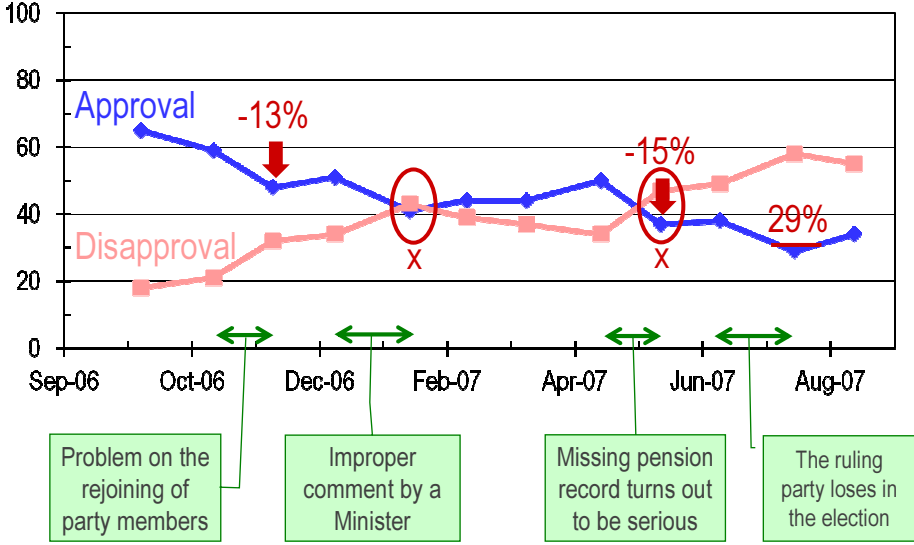


Fig. 2. An example of cabinet approval/disapproval rates and the detection of event periods in the case of ex-Prime Minister Shinzo ABE. When a dynamic behavior is observed in the graph, we consider that an important event occurred during the preceding period. The explanation in boxes are manually annotated to show the actual events that occurred. As a matter of fact, these events were mentioned in the corresponding Wikipedia article as causes of his resignation.

Topic-Based Event Detection. Since there are various events that could not be handled by the template-based approach, we decided to refer to the behavior of demoscopic polls. We referred to demoscopic polls provided monthly by NHK Broadcasting Culture Research Institute[4].

As shown in Fig. 2, we set the following conditions as drastic poll behaviors:

- Drastic increase / decrease in the approval rate (\uparrow/\downarrow).
- Reversal of approval and disapproval rates (X).
- Extremely high / low approval rate (—).

When either of the above conditions is observed, we considered that a major event occurred in the period of the current and the previous polls.

Next, news stories that describe the event during the period selected above are detected. In order to do so, we considered that the story should be either the beginning or the end of a news topic, or a heavily discussed story. We decided to measure such features from the topic thread structures proposed in [2]; A story is considered as a candidate that describes the major event if:

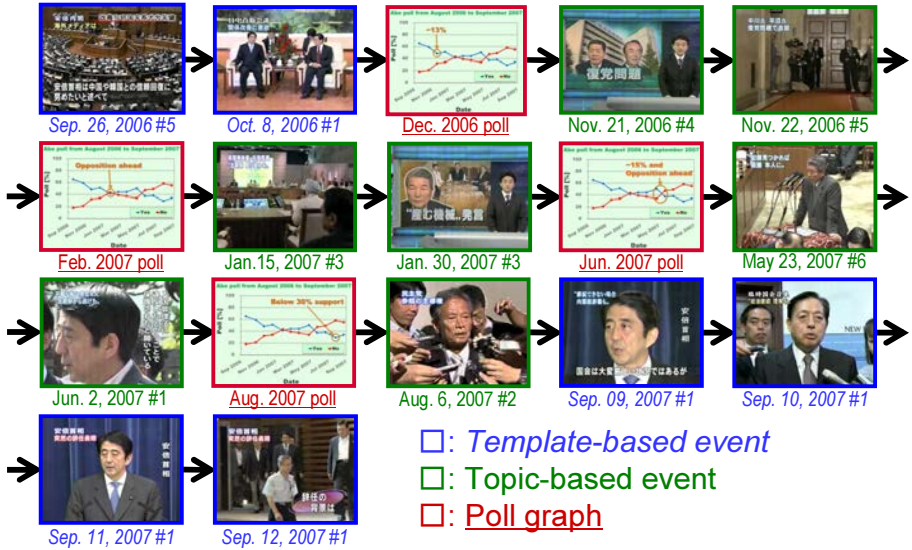


Fig. 3. Example of a generated summarized video in the case of ex-Prime Minister Shinzo ABE. Poll graphs are inserted when a drastic behavior is observed in the graph, followed by video clips that is supposed to contain descriptions on explanatory events (topic-based events). In the inauguration / resignation period, video clips are selected according to templates (template-based events).

- it is the beginning or the end of a topic thread structure, or
- stories are dense within a short period of time before and after it.

Finally, we apply sentiment analysis to the detected candidate stories referring to the dictionary created by Takamura et al. [5]. According to the user’s preference, the candidates during a period are ranked in optimistic or pessimistic order, and selected from the most extreme ones.

Editing. In order to produce a summarized video with a specific length, we need to select a certain number of video clips with a certain length. We cropped a certain length of video segment starting from the utterance of the Prime Minister’s name.

For the final output, poll graphs are inserted when a drastic behavior is observed in the graph, followed by video clips that is supposed to contain descriptions on explanatory events (topic-based events). In the inauguration / resignation period, video clips are selected according to templates (template-based events).

Fig. 3 shows an example of a summarized video produced by the proposed method.

3 Application to Sports Contents

For sports contents, we will introduce a work on video recompilation that makes use of the frequency and the estimated user attributes of twitter tweets as a “scenario”.

3.1 Biased Sports Video Summarization According to Twitter Tweets

- Video: Sports show
- Scenario: Twitter tweets (frequency and estimated user attributes)

Recently, micro-blogging services such as twitter has become very popular. Especially in events such as sports games, it has become common for thousands of people to tweet while watching the game, sharing the experience in real time. Following this trend, we proposed a framework that analyzes tweets concerning a sports game (i.e. those with hashtags related to the game) to produce a summarized video biased towards supporters of each team, and developed an interface to demonstrate the results called “twiSpo” (Fig. 4). Details of the method could be found in [6].

Estimation of User Attributes. First, the attribute of each user who tweeted with a hashtag related to the game is estimated. Since it is difficult to analyze the attribute (i.e. which team the user supports) of each user from a single tweet, it is analyzed from a sequence of the user’s tweets.

The classification was done by training keywords characteristic for tweets by users supporting each team. A dictionary for characteristic words was constructed by a method called SO-PMI (Semantic Orientation using Pointwise Mutual Information) [7].

Detection of Biased High-Light Scenes. Next, in order to detect high-light scenes biased towards each team, frequencies of tweets by supporters of each team are separately counted. Fig. 5 shows an example in the case of a baseball game. Based on thresholding to the frequency, summarized videos biased towards each team are produced.

4 Application to Cooking Contents

For cooking contents, we will introduce a work on video recompilation that makes use of text-based cooking recipes as a “scenario”.



Fig. 4. The “twiSpo” interface. From the console on the top-right of the interface, users can select the team which he/she supports, and also the number and the duration of high-light scenes to be included in the summarized video. At the bottom, tweets are also shown along the summarized video played on the top-left.

4.1 Description of Text Recipes with Videos

- Video: Cook show
- Scenario: Cooking recipe

Recently, cooking recipe sites that allow posting from general users have become popular, and hundreds of recipes are posted to such sites on a daily basis. Although it is easy to post in text to such sites, it is still infrequent to post images and moreover videos that describe the cooking procedures due to the complexity of the editing.

Since cooking procedures are sometimes difficult to understand without visual description, we proposed a framework that automatically links corresponding video clips in a database to cooking operations in an arbitrary text recipe, and developed an interface called “Video CooKing” as shown in Fig. 6 to demonstrate the results. Details of the method could be found in [8,9].

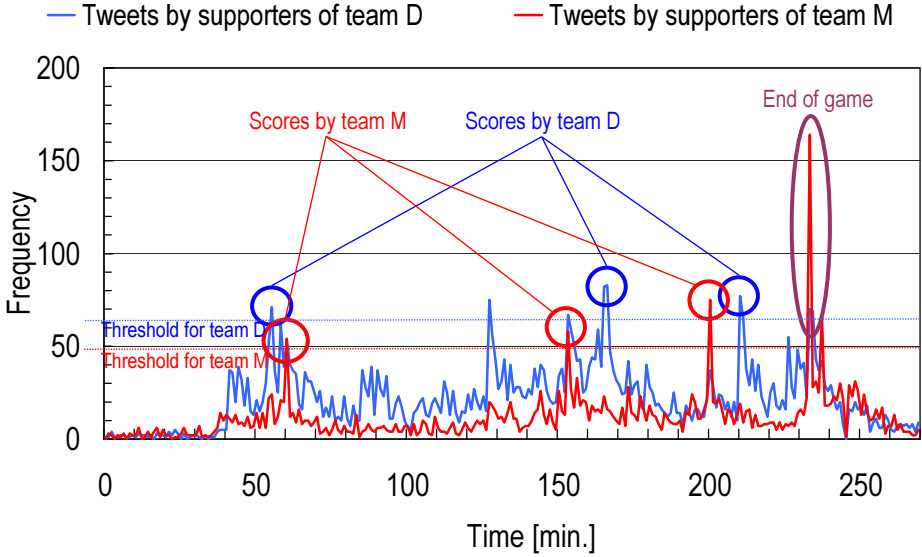


Fig. 5. Example of biased high-light scene detection. The two line graphs represent frequencies of tweets by supporters of each team. According to thresholds, high-light scenes are detected for each team. See the different high-light scenes detected.


Creation of a Video Database on Cooking Operations. In order to realize the framework, the most important part is the creation of a database that consists of video clips describing cooking operations. Since cooking operations may be different according to the ingredient to be cooked, the video clips should be annotated with a paired tag: (ingredient, operation).

Such a database could be created manually if possible, but due to the explosive number of combinations of ingredients and operations, we proposed a method to automatically create the database from cook shows broadcasted on TV. Most cook shows broadcasted in Japan come with closed-caption (audio transcript), so we analyzed the modification structure concerning ingredients that appear in it, in order to obtain (ingredient, operation) pairs that could be candidates for annotations.

However, the appearance in closed-caption does not always guarantee the existence of the actual cooking operation in the video. In addition, the video usually contains meaningless motions before or after the corresponding cooking operation. In order to correctly extract a video clip that describes a cooking operation corresponding to the (ingredient, operation) pair that appeared in the closed-caption around the same timing, motion features are analyzed.

As shown in Fig. 7, first, the motion in a video clip is classified into “repetitive”, “static”, or “others” by the trajectory of motion in the feature space. Next, repetitive motions are further classified in two by the distribution of repetitious motions in the frame. The motion class of the video clip is then matched with the

[Country-style vegetable soup]



Energy: 70kcal Cooking time: 20 min.
 Tag: Potato, Onion, Carrot, Cabbage, Soup, Bacon
 Source: http://www.kyounoryouri.jp/recipe/4162_田舎風野菜スープ.html

Ingredients (Serves 2)

- 1 cabbage leaf
- 1 1/4 carrot
- 1 1/2 onion
- 1 potato
- 1 slice of bacon
- a pinch of herb

Directions

1. Shred the cabbage leaf, cut in quarter-rounds the carrot. Slice the onion. Cut the potato half and slice them, soak them in water, exchange the water a few times, and then drain. Slice the bacon.
2. Put (1) in a heat-resistant container and add and mix salad oil, cover the container and heat it in a microwave.
3. Add salt, pepper, herb, and water into (2), and heat it in a microwave.
4. Take it out and serve it on a dish.

Visual explanations

Cooking operation: "Cut in quarter-rounds"

Ingredient: Carrot / Carrot(2)




Fig. 6. The “Video CookiNg” interface. The left side shows the original text recipe with links under cooking operations added by the proposed method. The right side shows video clips retrieved from the database which corresponds to the cooking operation specified in the text. There may be multiple corresponding video clips, so the user can learn different ways to perform the operation.

annotation candidates. In the end, if the classification matches, the annotation candidate is selected as the annotation for the video clip.

Linking Video Clips to a Text Recipe. When a text recipe is given, the modification structure concerning ingredients that appear in it is analyzed, in order to obtain (ingredient, operation) pairs. Next, the (ingredient, operation) pairs are sent to the video database as a query, and if available, corresponding video clips are linked from the text recipe. In case if there is no video clip corresponding to a certain (ingredient, operation) pair in the database, a partial match with only the operation is allowed.

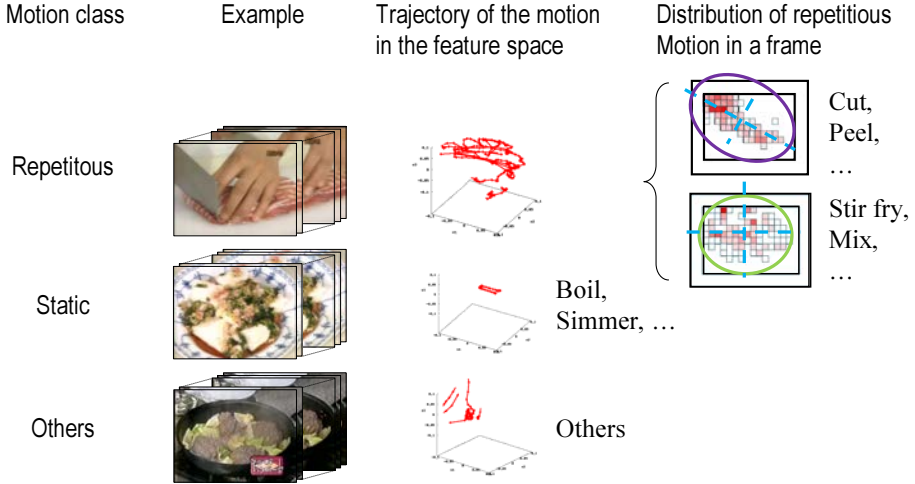


Fig. 7. Classification of cooking operations based on motion features. A video clip is classified first according to the trajectory of motion in a feature space, and then repetitive motions are further classified according to the distribution of motion in the frame.

5 Conclusion

In this paper, we introduced our works in news, sports, and cooking domains, that makes use of Wikipedia articles, demoscopic polls, twitter tweets, and cooking recipes in order to recompile video clips from corresponding TV shows.

In order to create contents bearable for practical use, besides improving the performance of individual techniques, we will need to collect and incorporate materials with a higher variety. We are considering to do so by establishing a framework that also makes use of videos available on the Web.

Acknowledgement. Parts of the works introduced in this paper were supported by Grants-in-aid for Scientific Research (B), for Scientific Research on Priority Areas (Infoplosion), and for Young Researchers (B), together with JSPS Excellent Young Researcher Overseas Visit Program, and joint research projects with National Institute of Informatics, Japan, and University of Amsterdam, the Netherlands. We would also like to thank the faculty and students involved in each work.

References

1. Okuoka, T., Takahashi, T., Deguchi, D., Ide, I., Murase, H.: Labeling news topic threads with Wikipedia entries. In: Proc. 11th IEEE Int. Symposium on Multimedia, pp. 501–504 (2009)

2. Ide, I., Kinoshita, T., Takahashi, T., Mo, H., Katayama, N., Satoh, S., Murase, H.: Efficient tracking of news topics based on chronological semantic structures in a large-scale news video archive. *IEICE Trans. Information and Systems* E95-D(5), 1288–1300 (2012)
3. Nack, F., Ide, I.: Why did the Prime Minister resign? —Generation of event explanation from large news repositories—. In: *Proc. 19th ACM Int. Multimedia Conf.*, pp. 313–322 (2011)
4. NHK Broadcasting Culture Research Institute: *The NHK Monthly Report on Broadcast Research*. NHK Publishing, Inc., ISSN: 0288-0008
5. Takamura, H., Inui, T., Okumura, M.: Extracting semantic orientations of words using spin model. In: *Proc. 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 133–140 (2005)
6. Kobayashi, T., Noda, M., Deguchi, D., Takahashi, T., Ide, I., Murase, H.: Summarizing sports video by on-the-spot comments on twitter (in Japanese). *IEICE Technical Report*, MVE2010-162 (2011)
7. Turney, P.D.: Thumbs up? Thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: *Proc. 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–424 (2002)
8. Doman, K., Kuai, C.Y., Takahashi, T., Ide, I., Murase, H.: Video CooKing: Towards the Synthesis of Multimedia Cooking Recipes. In: Lee, K.-T., Tsai, W.-H., Liao, H.-Y.M., Chen, T., Hsieh, J.-W., Tseng, C.-C. (eds.) *MMM 2011 Part II. LNCS*, vol. 6524, pp. 135–145. Springer, Heidelberg (2011)
9. Doman, K., Kuai, C.-Y., Takahashi, T., Ide, I., Murase, H.: Smart Video CooKing: A multimedia cooking recipe browsing application on portable devices. In: *Proc. 20th ACM Int. Multimedia Conf.* (to appear, 2012)

Evaluating Novice and Expert Users on Handheld Video Retrieval Systems

David Scott, Frank Hopfgartner, Jinlin Guo, and Cathal Gurrin

Dublin City University

Glasnevin

Dublin 9, Ireland

{dscott, fhopfgartner, jguo, cgurrin}@computing.dcu.ie

Abstract. Content-based video retrieval systems have been widely associated with desktop environments that are largely complex in nature, targeting expert users and often require complex queries. Due to this complexity, interaction with these systems can be a challenge for regular "novice" users. In recent years, a shift can be observed from this traditional desktop environment to that of handheld devices, which requires a different approach to interacting with the user. In this paper, we evaluate the performance of a handheld content-based video retrieval system on both expert and novice users. We show that with this type of device, a simple and intuitive interface, which incorporates the principles of content-based systems, though hidden from the user, attains the same accuracy for both novice and desktop users when faced with complex information retrieval tasks. We describe an experiment which utilises the Apple iPad as our handheld medium in which both a group of experts and novice users run the interactive experiments from the 2010 TRECVID Known-Item Search task. The results indicate that a carefully defined interface can equalise the performance of both novice and expert users.

Keywords: Mobile Device, Keyframe, iPad.

1 Introduction

There has been an evident shift in the way we access online content, with the advent of handheld devices and smart phones we have moved away from the rigid structured nature of desktop and laptop environments and embraced the portability and ease-of-use of mobile devices. The level of ubiquitous, always on access that handheld devices provide results in the average user having access to a WWW of information and a small device with (still) limited interaction capabilities to access it. This rush to handheld is further backed by a recent survey carried out by Morgan Stanley¹, estimating that by 2013 handheld devices will have overtaken desktop systems as the most popular portal to the web. It is now apparent that people are likely to access the web from handheld devices in

¹ <http://www.morganstanley.com/institutional/techresearch/>

a variety of environments, resulting in a search experience that is significantly more cognitively challenging than it was the case a few years ago when we could assume that a user was accessing a video search engine from a desktop computer.

There has been a lot of research efforts in recent years on the development of video search engines using a myriad of available computing devices [6]. A lot of this research has been undertaken through activities in conferences such as TRECVID² and VideoCLEF³. These video benchmarking conferences encourage knowledge sharing and publication of video search techniques and support the cross-site evaluation of state-of-the-art systems. Participation in conferences such as TRECVID is open worldwide with participants such as Carnegie Mellon University with their Informedia system [15] and University of Amsterdam with their MediaMill system [13] developing novel systems in recent years. While most of this type of research has focused on desktop interaction, TRECVID participants have recently begun to address the new handheld technologies in their video search engine evaluations. For example, DCU's TRECVID submission in 2010 utilised an iPad interface [2] and evaluated the effectiveness of this iPad video search engine on both novel and experienced video search users.

To this end we have focused mainly on content-based video retrieval and the development of new search techniques that can support mobile device access to digital video archives. We want to keep processing on the mobile device to a minimum and not burden the user with excessive requirements for complex interaction with on-screen elements or detailed examination of result sets to identify if the desired information is contained in the particular video document. We strive to develop a simple interface utilising previous knowledge such as storyboarding of video keyframes, utilising of concepts and similarity search [3,9,14] to provide expert searchers with the familiarity of traditional content-based systems while introducing novice users to a new and novel way to search, by hiding the complexity of content-based retrieval operation.

In the remainder of this paper we will describe the video retrieval system used by both our novice and expert users. Following this, we will outline our experiments and discuss the results attained by both NIST and by analysis of the post experiment user logs. Finally, we will draw our conclusions and present possible future work.

2 System Overview

Our system was developed as a native iPad application, incorporating a front-end interface and a back-end web service connected to databases and search engine indexes. This provides three methods of searching to facilitate both our expert and novice users: two primary methods (free-text and context search) and one secondary method (similarity search):

² <http://trecvid.nist.gov/>

³ <http://www.multimediaeval.org/>

- **Free Text Search:** The first of our primary search methods supports textual querying over three text indexes; meta-data, automatic speech recognition (ASR) text and a phonetically encoding text retrieval system.
- **Concept Search:** The second of our primary search methods, computer vision models trained via Support Vector Machines (SVM), provides a ranked list of the occurrence of any chosen concept e.g. Person, Vehicle, Computer Screen, Face, etc. from within the video archive.
- **Similarity Search:** Our secondary search method implements a relevance feedback technique that, for any given video document, returns a ranked list of the top fifty most visually similar keyframe representations within the collection.

2.1 Interface

From a user interface (UI) perspective, our goal was to develop a system that was easy and intuitive to use for both novice and expert users, while still allowing the user to utilise the available underlying search technologies. This trade-off between the power of functionality and simplicity of use is a well know design issue. By using an iPad device with a touch screen input and by developing a new interface specifically designed for that device we aimed to strike a balance between functionality and ease of use.

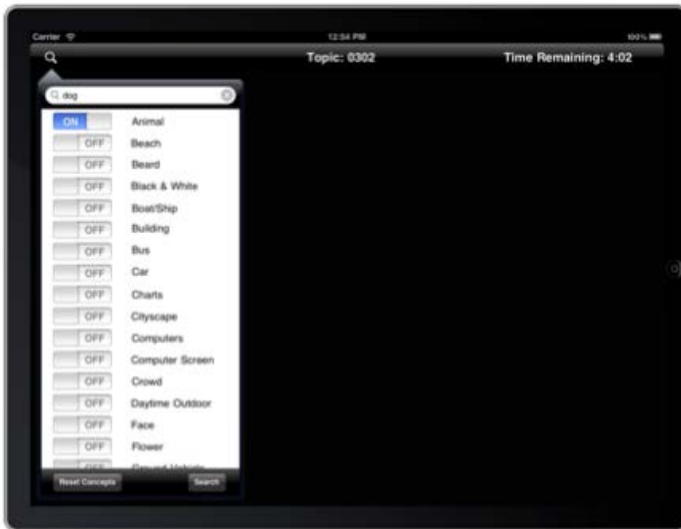


Fig. 1. Search Panel

Upon starting the application the user is required to enter a unique user ID, which allows the system to control the tasks assigned to the user and the system



Fig. 2. Search Results

can then track the progress of the user. Once the user has chosen to start a new topic they are presented with a search panel (as shown in Figure 1). Here, they can input a text query as well as select from a list of 33 predefined semantic concepts. The video results are returned in ranked order to the user: for each video, the title and description as well as a set of keyframes for each shot is shown (the user can scroll to the right to see more for each video). The top ranked shot for each video appears first in the list, with a maximum of 10 keyframes being displayed (selected temporally from throughout the video).

At any point during the search the user can tap on the search icon, which displays the search panel and allows them to refine their search. In addition to this, by double tapping on any keyframe the user can invoke a content-based image similarity search that returns video keyframes that appear visually similar to the one they have selected. After the allocated 5 minutes have elapsed or after the user successfully finds the relevant video the system returns to a topic start page.

2.2 Free Text Search

The text search engine we have chosen to use is Terrier developed by the University of Glasgow [12]. We created three separate indexes over the data and determined a weighting and fusion model by utilising test case topics and results as supplied by the TRECVID organisers.

Source Metadata: Contains metadata information pertaining to the video as crawled from the internet archive and supplied by the TRECVID organisers. The

information stored in this index includes author comments and are generally considered to provide a good overview of each video in the collection.

Automatic Speech Recognition: This index was created by utilising the spoken words in the video and was provided by LIMSI and Vecsys Research [5]. This information was indexed at the shot level by aligning the spoken word to the associated shot bounds.

Phonetic Encoding (PE): PE is concerned with representing the pronunciation of a word with a code made up of letters and numbers [1,8]. Similar words will have the same code and can therefore be matched by the search engine. Having performed an analysis of several techniques we found that the NYSIIS system [11] was the best choice for this experiment. The output of this process is a set of similar sounding words to the words in the meta-data and ASR which is then indexed by the search engine.

We chose this search engine structure as it increased recall, training topics had revealed that while the average rank fell from 250 to 700 with this index as opposed to a meta only index there was potential to discover 30% more known items with the three index setup.

2.3 Concept Search

Recent research has shown that systems based on the BoW model [7] produced the best results on several large scale content-based image and video retrieval benchmarks, we see a visualization of our approach outlined on Figure 3.

- SIFT Feature Extraction: The SIFT feature proposed by Lowe has proved to be very successful in applications such as object recognition and image retrieval. To compute SIFT features we use the version described by Lowe [10].
- Construction of Visual Vocabulary: In the construction of the visual vocabulary we employ the Hierarchical K-means algorithm to construct the visual vocabulary based on its advantages of simple and fast implementation. Five million SIFT descriptors were extracted from keyframes from the training data and these were clustered hierarchically using K-means to generate a vocabulary tree with 1296 leaf nodes (i.e. 1296 visual words).
- Visual Vocabulary Transformation: Soft assignment is utilised in the step of visual vocabulary transformation. For each key-point in an image, instead of mapping it only to its nearest visual word, in soft assignment we select the top-100 nearest visual words.

From here we use both positive and negative examples to feed into a Support Vector Machine (see Figure 3) to train the concepts, in the final system we developed 33 concepts based on types of concepts used in the training topics. They are: animal, beach, beard, black and white video, boat/ship, building, bus, car, charts, cityscape, computers, computer screen, crowd, daytime outdoor, face, flower, ground vehicle, in-door, indoor sports, landscape ,map, meeting, military, nighttime, office, outdoor, person, road, sky, snow, stadium, tree, and vegetarian.

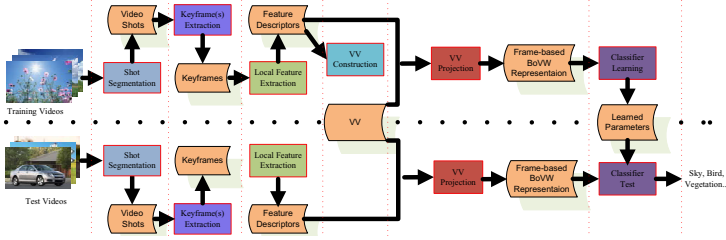


Fig. 3. Concept Training

2.4 Similarity Search

Content-based keyframe search allows users to select a shot on the interface and to find shots visually similar from the collection. For each keyframe in the collection we extracted three low-level MPEG-7 features, namely Colour Layout, Edge Histogram and Scalable Colour.

For each feature we calculated the similarity between each pair of keyframes in the collection. In order to reduce the space requirement for storing the resulting indexes we only stored the top 1,000 similar keyframes for each keyframe. Having calculated the set of similar keyframes for each keyframe in the collection we then combine the scores for each feature into an overall similarity score for a pair of keyframes. For data fusion we first normalise using MinMax normalisation, using CombSUM[4] to combine the normalised result lists.

3 Experiment

Through our experiments we wanted to compare the performance of novice users against expert users when using a feature-rich content-based retrieval system for video data. In particular we wanted to see if we could develop a single search system which could be used by novices and experts alike, with equal performance being attained by both. In addition, we were interested to compare the performance of our iPad search system against other systems taking part in the TRECVID 2010 evaluations. We recruited six users from our research group to complete the task in-house. All of these users had experience working with content-based video search systems and many had participated as users in previous TRECVID experiments completed in DCU, as such this group represents our expert users. We also recruited 12 users to participate from the BI School of Management in Oslo, Norway. None of these users had experience using a sophisticated content-based video search system and none had hands-on experience with using an iPad before. These users represent our novice users. Each participant completed one training topic, followed by 12 search topics during the experiments. We used the Latin-squares experimental design in order to assign users to topics and the ordering of presenting topics to each user was randomised in order to reduce the effects of learning bias, see Table 1.

Table 1. Table outlining the topic distribution between the novice and expert user groups

Novice:	1	2	3	4	5	6	7	8	9	10	11	12
Expert:	1	2	3	4	5	6	7	8	9	10	11	12
Topic 1:	x	x	x				x	x	x			
Topic 2:	x	x	x				x	x	x			
Topic 3:	x	x	x				x	x	x			
Topic 4:	x	x	x				x	x	x			
Topic 5:	x	x	x				x	x	x			
Topic 6:	x	x	x				x	x	x			
Topic 7:	x			x	x		x			x	x	
Topic 8:	x			x	x		x			x	x	
Topic 9:	x			x	x		x			x	x	
Topic 10:	x			x	x		x			x	x	
Topic 11:	x			x	x		x			x	x	
Topic 12:	x			x	x		x			x	x	
Topic 13:		x	x	x		x	x	x		x	x	x
Topic 14:		x	x	x		x	x	x		x	x	x
Topic 15:		x	x	x		x	x	x		x	x	x
Topic 16:		x	x	x		x	x	x		x	x	x
Topic 17:		x	x	x		x	x	x		x	x	x
Topic 18:		x	x	x		x	x	x		x	x	x
Topic 19:			x		x	x			x		x	x
Topic 20:			x		x	x			x		x	x
Topic 21:			x		x	x			x		x	x
Topic 22:			x		x	x			x		x	x
Topic 23:			x		x	x			x		x	x
Topic 24:			x		x	x			x		x	x

The interactive known-item search task at TRECVID 2010 had six teams submit a total of 14 runs. Each run belonged to a certain category depending on the training type and whether the meta-data XML was used or not. For both of our runs we used the meta-data XML (condition: YES) and used only the IACC training data (training type: A). Each system was evaluated based on Mean Elapsed Time, an average of the times recorded for each topic with topics not found being assigned the maximum five minutes. Figure 4 presents the results for all submissions to the interactive known-item search, our two runs are highlighted. Both runs represent results from multiple users where we have picked the best time for each topic in order to populate our submission. Overall our runs came 6th and 7th, however when we compare ourselves against groups with the same condition and training type the position is 5th and 6th.

In the expert run there were a total of 9 topics (out of a total of 22) for which none of our participants found the correct video, interestingly the novice users only missed 8. The fact that users could not find the correct video for certain topics is not surprising, having observed the user experiments it was clear that users found the majority of topics to be either very easy or very difficult. Perhaps more interestingly for us, as part of our post-experiment questionnaire we asked our users to score the system in terms of ease-of-use on a scale of 1–7. For this, our novice users gave the system a median score of 6, with experts giving a median score of 6.5.

Post experiment analysis showed that our novice users executed more searches than our expert users (64 vs 53 on average). From the chart in Figure 5 we see the difference between the users utilisation of the search features; the novice users appeared more reluctant to use both the concept only search (0.88% of the time) and similarity search (5.72% of the time), thus resulting in their requiring on average 9 more queries per experiment (all twelve topics) per experiment to

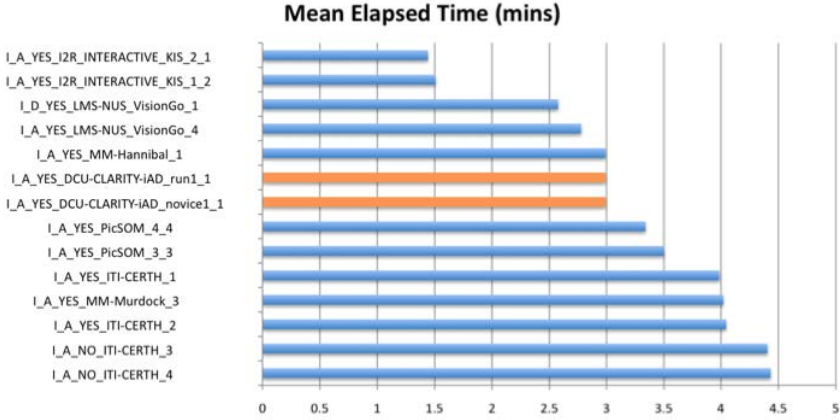


Fig. 4. Official TRECVID Results

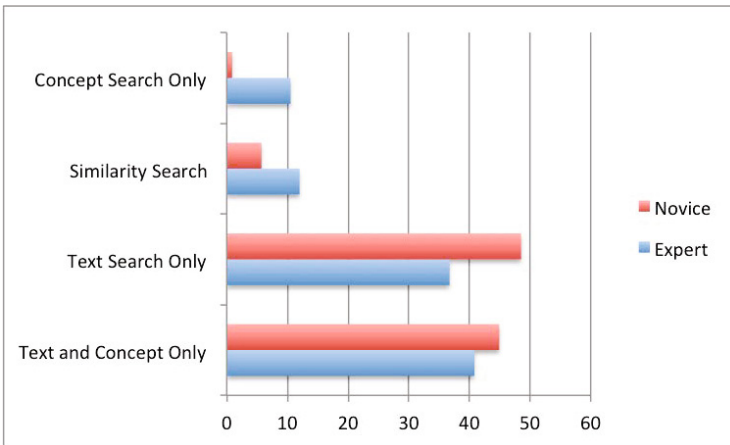


Fig. 5. Log Analysis

attain the same results as our experts. We see from the graph that novice users’ preferred search method is text only search nearly 50% of all searches as opposed to experts’ who preferred both text and concept search.

4 Conclusions and Future Work

In this work, we developed a video search system aimed at integrating complex search techniques into a single easy-to-use handheld video search engine that

can be used equally as effectively by experts and novice users. The results from our official experiments show that the performance of novices versus experts is identical in terms of mean elapsed time. Through our post-experiment analysis we are investigating why this is the case. One explanation would be that our attempts to build a search engine that could be used by novices and experts alike was successful. Another explanation could lie in the topics used in the search task. Through observations of the experiments we found that both sets of users found the majority of topics to be either very easy or very difficult. The lack of topics of medium difficulty may have constrained our ability to distinguish the differences in performance of different users. Nonetheless through our experimental logs and questionnaires we can still gain valuable insights into the techniques used by both sets of users and their experiences in using our system.

Having analysed the log files further we noted that while the novice users and the experts got the same results according to the official results they relied heavily on text based searches, we intend to further aid these users by incorporating the visual features by using clustering techniques to group like keyframes, this will allow users to more quickly identify relevant keyframes and dismiss a group at a glance thus speeding up their Mean Elapsed Time.

Acknowledgments. The research was funded by Information Access Disruptions, a centre for research-based innovation with CRI number: 174867, funded in part by the Norwegian Research Council.

References

1. Elmagarmid, A.K., Ipeirotis, P.G., VerykiosDuplicate, V.S.: record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering* 19, 1–16 (2007)
2. Foley, C., Guo, J., Scott, D., Wilkins, P., Gurrin, C., Smeaton, A.F., Ferguson, P., Cusker, K.M., Diaz, E.S., McGuinness, K., O'Connor, N.E.: TRECVID 2010 Experiments at Dublin City University. In: *Proceedings of the 8th TRECVID Workshop*, Gaithersburg, USA (November 2010)
3. Foley, C., Gurrin, C., Jones, G., Lee, H., Givney, S.M., O'Connor, N., Sav, S., Smeaton, A.F., Wilkins, P.: TRECVID 2005 Experiments at Dublin City University. In: *TRECVID 2005 - Text REtrieval Conference TRECVID Workshop*, MD, USA, National Institute of Standards and Technology (2005)
4. Fox, E.A., Shaw, J.A.: Combination of Multiple Searches. In: *Text REtrieval Conference*, pp. 243–249 (1994)
5. Gauvain, J.-L., Lamel, L., Adda, G.: The LIMSI Broadcast News transcription system. *Speech Commun.* 37(1-2), 89–108 (2002)
6. Hopfgartner, F.: *Understanding Video Retrieval*. VDM Verlag (2007)
7. Jiang, Y.-G., Yang, J., Ngo, C.-W., Hauptmann, A.G.: Representations of keypoint-based semantic concept detection: A comprehensive study. *Trans. Multi.* 12(1), 42–53 (2010)

8. Khan, A.M., Mckinley, K.S., Bentzur, R., Feinberg, D., Frampton, D., Guyer, S.Z., Hirzel, M., Hosking, A., Jump, M., Lee, H., Eliot, J., Moss, B., Phansalkar, A., Stefanovic, D., Vandrunen, T., Von Dincklage, D., Christen, P., Christen, P.: A comparison of personal name matching: Techniques and practical issues. In: Workshop on Mining Complex Data (MCD 2006), held at IEEE ICDM 2006, Hong Kong, pp. 290–294 (2006)
9. Koskela, M., Wilkins, P., Adamek, T., Smeaton, A.F., O'Connor, N.: TRECVID 2006 Experiments at Dublin City University. In: TRECVID 2006 - Text REtrieval Conference TRECVID Workshop (2006)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int'l J. Computer Vision* 60, 91–110 (2004)
11. NYSIIS. Comprehensive perl archive network, <http://search.cpan.org/?krburton/String-Nysiis-1.00/Nysiis.pm>
12. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: Proceedings of ACM SIGIR 2006 Workshop on Open Source Information Retrieval, OSIR 2006 (2006)
13. Snoek, C.G.M., van de Sande, K.E.A., de Rooij, O., Huurnink, B., Gavves, E., Odijk, D., de Rijke, M., Gevers, T., Worring, M., Koelma, D.C., Smeulders, A.W.M.: The MediaMill TRECVID 2010 semantic video search engine. In: Proceedings of the 8th TRECVID Workshop, Gaithersburg, USA (November 2010)
14. Wilkins, P., Adamek, T., Jones, G., O'Connor, N., Smeaton, A.F.: TRECVID 2007 Experiments at Dublin City University. In: TRECVID 2007 - Text REtrieval Conference TRECVID Workshop (2007)
15. Chen, M.Y., Li, H., Hauptmann, A.: Informedia @ TRECVID 2009: Analyzing Video Motions. In: Proceedings of the 7th TRECVID Workshop, Gaithersburg, USA (November 2009)

Perfect Snapping

Qingsong Zhu^{1,2,3,4}, Ling Shao⁵, Qi Li^{1,6}, and Yaoqin Xie^{1,2,3,4,*}

¹ Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences,
518055, Shenzhen, China

² Key Lab for Low-Cost Healthcare, Chinese Academy of Sciences,
518055, Shenzhen, China

³ Key Lab for Health Informatics, Chinese Academy of Sciences,
518055, Shenzhen, China

⁴ School of Medicine, Stanford University, Stanford, California, USA

⁵ Department of Electronic and Electrical Engineering, University of Sheffield,
Sheffield, S10 2TN, UK

⁶ School of Software Engineering, University of Science and Technology of China,
230026, Hefei, China
qs.zhu@siat.ac.cn

Abstract. Interactive image matting is a process that extracts a foreground object from an image based on limited user input. In this paper, we propose a novel interactive image matting algorithm named Perfect Snapping which is inspired by the presented method named Lazy Snapping technique. In the algorithm, the mean shift algorithm with a boundary confidence prior is introduced to efficiently pre-segment the original image into homogeneous regions (super-pixels) with precise boundary. Secondly, Gaussian Mixture Model (GMM) clustering algorithm is used to describe and to model the super-pixels. Finally, a Monte Carlo based Expectation Maximization (EM) algorithm is used to perform parametric learning of mixture model for priori knowledge. Experimental results indicate that the proposed algorithm can achieve higher matting quality with higher efficiency.

Keywords: Interactive Image Matting, Mean Shift Algorithm, Lazy Snapping.

1 Introduction

Interactive image matting has been an important technique in image processing and video editing, which refers to the problem of softly extracting the foreground objects from a single input image. With the rapid development of digital image processing techniques, image matting has become possible to segment the foreground objects on an individual pixel level. And a variety of image matting algorithms have been proposed and used in post image and video editing. Purpose of image matting is to specify which parts of the image belong to the foreground and which parts belong to background. Usually, a user imposes certain hard constraints for

* Corresponding author.

segmentation by indicating certain pixels (seeds) that absolutely have to be part of the foreground and certain pixels that have to be part of the background. Intuitively, these hard constraints provide clues on what the user intends to segment.

The early matting algorithms are based on known background. The blue screen matting [1] was used for live action matting, whose principle is to photograph the subject against a constant-colored background (typically blue and green). Recently, in the field of image matting study, many natural image matting approaches have been proposed. In natural image matting processing, moderate user interactions are essential. In the Knockout [2] method, the algorithm starts from the known foreground and background of the trimap and extrapolates the known foreground and background colors into the unknown region to estimate the alpha matte. In Ruzon and Tomasi's approach [3], a statistical method is proposed to analyze the color samples of the foreground and background and the estimation of α . In [4], a new image matting algorithm based on principal components analysis (PCA) is introduced to analyze the foreground and background samples. This method utilizes the projection method to estimate α and has also a considerable computation cost. In [5], a Bayesian framework based image matting approach is proposed. In this method, color samples of foreground and background are clustered and modeled as mixture Gaussian distribution. An estimation named maximum a posterior (MAP) is introduced to calculate foreground, background and α simultaneously for each pair of the foreground and background in a Bayesian framework. The final estimate of α is chosen from the pair of the foreground and background that provides the maximum likelihood. Although the algorithm achieved a better result of matting, the algorithm is quite complex and has a slower processing speed compared with Knockout method. In [6], the Poisson Matting approach for natural images matting of complex scenes is presented. The basic idea behind the algorithm is to formulate the problem of natural image matting as one of solving Poisson equations with the matte gradient field. Unlike previous methods that optimize a pixel's alpha matte in a statistical manner, the Poisson Matting method operates directly on the gradient of the matte, which reduces greatly the error caused by mis-classification of color samples in a complex scene. But the shortcomings of the algorithm include two aspects. First, when the foreground and background colors are very similar, the matting equation becomes ill-conditioned. Secondly, when the matte gradient estimated in global Poisson Matting largely biases the true values, more user interaction is required.

In [7], a novel approach called Flash Matting is proposed. This algorithm can robustly recover the matte from flash/no-flash images, even for scenes in which the foreground and the background are similar or the background is complex. In [8], an image matting approach based on Belief Propagation is presented. The approach can achieve better matting results only with less trimap restriction by utilizing the discrete set of α value to formulate the matting problem as energy minimum problem. But the algorithm is quite time consumed subject to the iterative processing. Another scribble-based method is proposed in [9]. The algorithm has a better interaction performance and can achieve a better result

only by a few scribbles restriction. But it has a higher computation cost and a lack of color statistic characteristic. A known trimap is essential for the above image matting algorithms. In [10-13], some interactive Graph Cuts [10] based approaches are introduced. Grab Cut method [11] makes user draw a rectangle around the periphery of the foreground object, and then extracts the foreground objects accurately by image segmentation and feathered process. Grow Cut approach [12] is a geodesic distance based image matting algorithm which utilizes the surface of “hard constraint pixels” that user calibrated to grow outside to complete image matting, and it is difficult to operate the texture region using the method. Lazy Snapping [13] is another well-known image matting algorithm. User only draws a few lines in different places. The region that some lines appear is regarded as foreground region and the region that others lines appear is regarded as background region, and user can separate the object from the background by these lines. However, the imposed lines drawn by user must satisfy sufficiently to represent the colors species in the foreground and background region. Otherwise user must constantly add new lines until get satisfactory results. Moreover, the over-segmentation problem of the method remains to be solved.

In this paper, a novel image matting algorithm named Perfect Snapping is proposed. The algorithm can be divided into following steps: a) Use mean shift algorithm with a boundary confidence prior to efficiently pre-segment the original image into homogeneous regions (super-pixels); b) Perform mainly description and modeling for the super-pixels by Gaussian Mixture Model clustering algorithm; c) Complete the parametric learning of mixture model for priori knowledge. Extensive experimental results have been implemented and compared with classical algorithm to show its advantage.

2 Perfect Snapping Algorithm

2.1 Graph Cut with Pre-segmentation

The main idea of the Graph Cut model is to construct an energy function and use weighted graph mapping and network flow theory to convert the global labeling problem to the maximum-flow/minimum cut problem of the corresponding weighted graph. To indicate the classification of each pixel, we can construct pixel-based Markov Random Field energy function as:

$$E(X) = \sum_{i \in \nu} E_1(x_i) + \lambda \sum_{(i,j) \in \varepsilon} E_2(x_i, x_j) \quad (1)$$

Where ν is the set of all nodes and ε is the set of all arcs connecting adjacent nodes. $E_1(x_i)$ is the likelihood energy which measures the energy consumption that a node i is defined as foreground or background, and $E_2(x_i, x_j)$ is the prior energy that denotes the cost when the labels of adjacent nodes i and j are x_i and x_j respectively. In order to simplify Graph Cut model, we usually use some pre-segmentation algorithms to segment the original image into some small regions which are regarded as the nodes of the weighted graph to construct the Graph

Cut model. Compared with traditional Graph Cut method viewing the pixel as node, the approach greatly simplify the topological structure of weighted graph and reduced the computation cost. In this paper we introduce a mean shift based pre-segment algorithm with boundary prior in place of the watershed method appeared in Lazy Snapping [12]. Mean shift algorithm is an efficient tool used for feature space analysis. To make segmentation results similar in color and continuous in space, we perform a mean shift filtering on an original image in 5-D feature space $l^*u^*v^* \sim x^*y^*$. Assume that the probability density function of 5-D feature space is $f(x)$:

$$\nabla f(x) \propto \sum_{i=1}^n (x - x_i) \nabla k \left(\|h^{-1}(x - x_i)\|^2 \right) \quad (2)$$

Where $x_i \in W_{h,z}$, $W_{h,z}$ represents 5-D super-spheroid with center at points x_i and has a $h = \{h_s, h_c\}$ bandwidth. h_s and h_c represent the bandwidth of space and color, respectively. Let the function $g(x) = -k'(x)$ and the corresponding new kernel $G(x) = \lambda g \cdot \|x\|^2$, then the density of new kernel is described by:

$$\nabla f'(x) \propto \sum_{i=1}^n (x - x_i) g(\|h^{-1}(x - x_i)\|^2) \quad (3)$$

To improve the filtering speed, the pixels are only relegated to the corresponding model attractive regions. After filtering, the model attractive regions are executed recursion and combination according to regions adjacency graph algorithm, color bandwidth and the size of region. To obtain accurate pre-segment results, the mean shift algorithm is extended to incorporate a boundary confidence prior. Suppose that the gradient of a continuous surface $\omega(x, y)$ at (x, y) is the vector pointing toward the direction of largest increase on the surface as

$$\nabla \hat{w}(x, y) = \frac{\partial w}{\partial x} \mathbf{i} + \frac{\partial w}{\partial y} \mathbf{j} \quad (4)$$

Any Cartesian $x - y$ coordinate system can be chosen since it is easy to verify that the gradient magnitude and an edge orientation as:

$$\hat{\omega} = \|\nabla \hat{w}(x, y)\| = \left[\left(\frac{\partial w}{\partial x} \right)^2 + \left(\frac{\partial w}{\partial y} \right)^2 \right]^{\frac{1}{2}} \quad (5)$$

$$\hat{\theta} = \arctan \left(\frac{\partial w}{\partial y} / \frac{\partial w}{\partial x} \right) \quad (6)$$

After finishing gradient estimation, every pixel in the image is associated with an edge magnitude $\hat{\omega}$ and an edge orientation $\hat{\theta}$. Let $\hat{\omega}_{[1]} < \dots < \hat{\omega}_{[k]} < \hat{\omega}_{[k+1]} < \dots < \hat{\omega}_{[N]}$ be the ordered set of distinct magnitudes values. Therefore, for any pixel, its edge magnitude $\hat{\omega}_{[k]}$ is replaced with the probability:

$$\delta_k = \text{prob}[\hat{\omega} \leq \hat{\omega}_{[k]}] \quad (7)$$

Note that δ_k is the percentile of the cumulative gradient magnitude distribution. While the weight of each pixel i is described as:

$$\psi_i = 1 - [\alpha_i \xi_i + (1 - \alpha_i) \zeta_i] \quad (8)$$

Where ξ_i and ζ_i represent the estimated gradient magnitude and the confidence in the presence of an edge pattern, respectively. The nearer the pixels to an edge, the smaller its weight is. The above process can pre-segment the original image into many small regions whose edges are described well and whose color is consistent. In this paper, we define this region as super-pixel and use it to construct Graph Cut model. Compared with traditional single-pixel based model, our method can simplify the number of nodes and weighted edges of weighted graph topological structure and reduce the computation cost and memory consumption.

2.2 Character Description and Clustering

To extract the feature information of super-pixel, usual methods are to compute the feature average of all sample points of the region, which leads to the lack of spatial color correlation between pixels. Thus, in this paper we introduce a Gaussian Mixture Model clustering algorithm to describe the super-pixel. Denote a super-pixel i by $s_i = \{\mu_i, \Sigma_i\}$, where μ_i and Σ_i represent mean and variance of color feature of region i respectively. In Equation (1), to compute $E_1(x_i)$, the user can define the background seeds and the background seeds, the super-pixels of unknown regions can be clustered by Gaussian Mixture Model. The mean colors of the foreground and background clusters are denoted as $\{G_1^F, G_2^F, \dots, G_M^F\}$ and $\{G_1^B, G_2^B, \dots, G_N^B\}$ respectively, where M and N are the clusters of foreground and background respectively. Then, for each super-pixel, minimum distance from its color cluster G_i to foreground clusters can be expressed as

$$D_i^F = \min_{n \in [1, M]} dis(G_i, G_n^F) \quad (9)$$

$$D_i^B = \min_{n \in [1, N]} dis(G_i, G_n^B) \quad (10)$$

Therefore $E_1(x_i)$ is defined as follows:

$$E_1(x_i = 1) = \infty \quad E_1(x_i = 0) = 0 \quad \forall i \in B \quad (11)$$

$$E_1(x_i = 1) = 0 \quad E_1(x_i = 0) = \infty \quad \forall i \in F \quad (12)$$

$$E_1(x_i = 1) = D_i^F (D_i^F + D_i^B)^{-1} \quad \forall i \in U \quad (13)$$

$$E_1(x_i = 0) = D_i^B (D_i^F + D_i^B)^{-1} \quad \forall i \in U \quad (14)$$

Here, U is the uncertain (not labeled) super-pixel set. The third equation guarantees the super-pixels to have the label with similar colors to foreground or background. We define $E_2(x_i, x_j)$ as a function of the color gradient between two super-pixels i and j :

$$E_2(x_i, x_j) = |x_i - x_j| \cdot \exp \{-\phi dis^2(G_m, G_n)\} \quad (15)$$

$$\phi = \left(|Z|^{-1} \cdot \sum_{m,n \in Z} dis^2(G_m, G_n) \right)^{-\varepsilon} \quad (16)$$

$$dis(G_m, G_n) = \frac{\sqrt{2}}{2} \cdot \sqrt{KLD(G_m \| G_n) + 2KLD(G_n \| G_m)} \quad (17)$$

$$KLD(G_m \| G_n) = \int G_m(x) \log \frac{G_m(x)}{G_n(x)} dx \quad (18)$$

Where $KLD(\cdot)$ is abbreviation to Kullback Leibler Divergence, which is used to measure quantitatively the distance between Gaussian features. To perform parametric learning of mixture model for interactive priori knowledge, EM algorithm is usually better selection. EM algorithm is suitable for maximum likelihood based Graph Cut segmentation model. To overcome the problem of slow convergence speed of traditional EM algorithm, a Monte Carlo based EM (MCEM) acceleration algorithm is introduced. The main idea is to combine MCEM algorithm and Newton-Raphson algorithm and use Monte Carlo simulation to realize E-step of EM algorithm, which can not only preserve the advantage of EM algorithm but also effectively improve the convergence of EM algorithm. Finally, the full description of MCEM algorithm can be given. Firstly (E-step), use $p(\theta|Y, Z)$ as the posterior distribution density function of θ with adding the data Z , let $Q(\theta|\theta^{(i)}, Y)$ be E-step integral, given sampling spots $\{z_1, z_2, \dots, z_m\}$ from $p(Z|\theta^{(i)}, Y)$ Computing:

$$\hat{Q}(\theta|\theta^{(i)}, Y) = \frac{1}{m} \cdot \sum_{j=1}^m \log p(\theta|z_j, Y) \quad (19)$$

Secondly (M-step), maximizing the function $\hat{Q}(\theta|\theta^{(i)}, Y)$ to work out $\theta_{EM}^{(i)}$ and satisfy (let $\Theta = \log p(\theta|Y)$):

$$\hat{Q}(\theta_{EM}^{(i)}|\theta^{(i)}, Y) = \max_{\theta} \hat{Q}(\theta|\theta^{(i)}, Y) \quad (20)$$

$$\theta^{(i+1)} = \theta^{(i)} + \left(- \frac{\partial^2 \Theta}{\partial \theta^2} \Big|_{\theta^{(i)}} \right) \left[\int \frac{\partial \Theta}{\partial \theta} p(z|y, \theta^{(i)}) dz \Big|_{\theta^{(i)}} \right] \left(\theta_{EM}^{(i)} - \theta^{(i)} \right) \quad (21)$$

Thus arose an iteration: $\theta^{(i)} \rightarrow \theta^{(i+1)}$ and then perform the iteration operation for the above E-step and M-step until $\|\theta^{(i+1)} - \theta^{(i)}\|$ or $\|Q(\theta^{(i+1)}|\theta^{(i)}, Y) - Q(\theta^{(i)}|\theta^{(i)}, Y)\|$ approaches infinitesimal. By the Geweke Law of Large Numbers, we have:

$$\hat{Q}(\theta|\theta^{(i)}, Y) \rightarrow Q(\theta|\theta^{(i)}, Y) \quad (22)$$

3 Experimental Result

To demonstrate the performance of our proposed approach, we first test it on some public images. We also compare our algorithm to Graph Cut [10], Grab Cut [11], Lazy Snapping [13]. Our algorithm starts with following initial parameters: $\lambda = 50, \alpha = 0.2, \varepsilon = 0.95, M = N = 5$. The system is running on a



Fig. 1. Pre-segmentation comparison of watershed and mean shift with a boundary confidence prior algorithm

P4-2GHz desktop with 1GB RAM. Figure 1 shows the pre-segmentation comparison of watershed appeared in Lazy Snapping [13] and mean shift used in our algorithm. The left column are the original test images, the 2nd and 3rd columns displayed the segmentation results by watershed and mean shift with a boundary confidence prior.

In comparison, the over-segmentation phenomenon of watershed is very serious, which necessarily leads to higher time complexity of sequent Graph Cut model. To compare, our method which uses mean shift incorporating a boundary confidence prior can effectively control the over-segmentation phenomenon and the number of regions pre-segmented is less than 1% of the watershed method.

The algorithm is also compared with *Graph Cut*, *Grab Cut* and *Lazy Snapping* and the results are as shown in Figure 2. In Figure 2, the top row is the original test images with a quick object marking step: the red lines are drawn to indicate the foreground and the blue lines to indicate the background. The 2nd, 3th, 4th and 5th rows displayed the matting results by Graph Cut, Grab Cut, Lazy Snapping and Our method respectively. In comparison, the proposed method outperforms in complex scenes (the extraction of a pair of thin and long tentacles in “Butterfly”, color similarity of foreground and background in “Fish”, background complexity in “Starfish” and “Boy”) and also gives better matting results compared with Graph Cut, Grab Cut and Lazy Snapping. In addition, comparison of execution time of the four methods is provided in Figure 3. By comparison, we can see that the time required of our method extracting foreground objects less than Graph Cut, Grab Cut and Lazy Snapping. For Lazy Snapping, in order to obtain a general satisfied result it need execute many interactions, and for any interaction operation it needs rerun the whole Graph Cut model and thus decreases its efficiency.

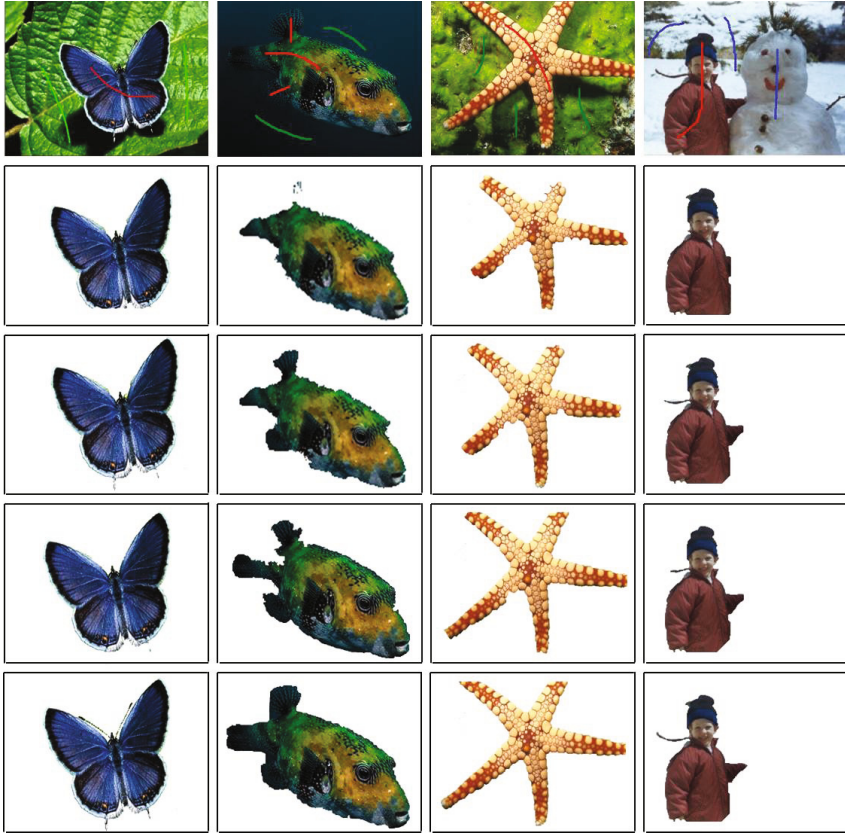


Fig. 2. Some comparative results by the four methods

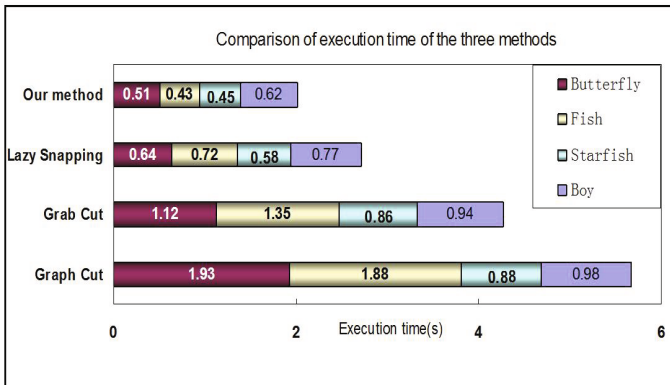


Fig. 3. Comparison of execution time of Graph Cut, Grab Cut, Lazy Snapping and the proposed method

4 Conclusion

In this paper, we propose a more effective interactive image matting method compared with Graph Cut, Grab Cut and Lazy Snapping. Firstly, our method uses mean shift algorithm with a boundary confidence prior to efficiently pre-segment the original image into homogeneous regions (super-pixels) with precise boundary. Secondly, we introduce Gaussian Mixture Model clustering algorithm to describe and model the super-pixels. Finally, a Monte Carlo based EM acceleration algorithm is presented to perform parametric learning of mixture model for priori knowledge. The experimental results show that our algorithm can outperform in both matting quality and efficiency.

References

1. Smith, A.R., Blinn, J.F.: Blue screen matting. In: Proceedings of SIGGRAPH, pp. 259–268 (1996)
2. Berman, A., Vlahos, P., Dadourian, A.: Comprehensive method for removing from an image the background surrounding a selected object. U. S patent, 134–345 (2000)
3. Ruzon, M., Tomasi, C.: Alpha estimation in natural images. In: Proc. CVPR, pp. 18–25 (2000)
4. Hillman, P., Hannah, J., Renshaw, D.: Alpha channel estimation in high resolution images and image sequences. In: Proc. CVPR, pp. 1063–1068 (2001)
5. Chuang, Y.Y., Curless, B., Salesin, D.: A Bayesian approach to digital matting. In: Proc. CVPR, pp. 264–271 (2001)
6. Sun, J., Jia, J., Tang, C.K., Shum, H.Y.: Poisson Matting. In: Proceedings of SIGGRAPH, pp. 315–321 (2004)
7. Sun, J., Li, Y., Kang, S.B., Shum, H.Y.: Flash Matting. In: Proceedings of SIGGRAPH, pp. 259–268 (2006)
8. Wang, J., Cohen, M.F.: An iterative optimization approach for unified image segmentation and matting. In: Proc. ICCV, pp. 936–943 (2005)
9. Levin, A., Lischinski, D., Weiss, Y.: A closed form solution to natural image matting. In: Proc. CVPR, pp. 61–68 (2006)
10. Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In: Proceedings of ICCV, pp. 105–112 (2001)
11. Rother, C., Blake, A., Kolmogorov, V.: Grab Cut-interactive foreground extraction using iterated graph cuts. In: Proceedings of SIGGRAPH, pp. 309–314 (2004)
12. Vezhnevets, V., Konouchine, V.: GrowCut- Interactive multi-label N-D image segmentation by cellular automata. In: Proc. Graphicon, pp. 150–156 (2005)
13. Li, Y., Sun, J., Tang, C.K., Shum, H.Y.: Lazy Snapping. In: Proceedings of SIGGRAPH, pp. 303–308 (2004)

Smart Video Browsing with Augmented Navigation Bars

Manfred Del Fabro, Bernd Münzer, and Laszlo Böszörményi

ITEC – Information Technology, Alpen-Adria-Universität Klagenfurt, Austria
{manfred,bernd,laszlo}@itec.aau.at

Abstract. While accuracy and speed get a lot of attention in video retrieval research, the investigation of interactive retrieval tools gets less attention and is often regarded as trivial. We want to show that even simple ideas have potential to improve the retrieval performance by giving some automated support to the browsing user. We present a video browsing concept where video segments are clustered in several latent classes of similar content. The navigation bars of our video browser are augmented with different colors indicating where elements of these clusters are located. As humans are able to classify the content of clusters fast, they can benefit from this information when browsing a video. We present a study where we investigated how humans can be supported in different video browsing tasks with a color-based and a motion-based clustering of video content.

1 Introduction

Video retrieval systems reached a good maturity level in recent years. In most cases the retrieval algorithms are trained to meet the characteristics of a specific use case or dataset. If they are applied to more general use cases or to different datasets, the accuracy typically degrades, making them impractical for real-world scenarios. Accurate automatic retrieval is of course important, but it is not the only aspect to consider when building a retrieval system. At the beginning and at the end of retrieval tasks there are always humans who must be served as good as possible in achieving their goals and intentions.

The human cognitive abilities are still much better than the abilities of automatic classifiers or matching algorithms and presumably they will ever be! At the *Multimedia Modeling Conference 2012*[22] the *Video Browser Showdown (VBS)* took place for the first time. Aim was to find the video browsing approach that is best suited for Known-Item-Search tasks (KIS). The winning tool completely refrains from content analysis[8]. Its success is only based on the combination of intelligent interaction means taking use of the human abilities to recognize and classify items fast.

From this experience we learn that a reasonable combination of automatic content analysis with human cognition might lead to the highest retrieval accuracy. Instead of trying to completely imitate the human cognitive abilities by automatic tools, we try to integrate humans as good as possible into the retrieval

process, without burdening too much work on them. The user interaction should be enhanced by automatic content analysis, but not replaced by it!

The presented idea can be seen as an analogy to augmented reality, where the real-world is augmented with additional information (e.g. with smartphone apps¹ or with special glasses²). People can take advantage of that additional information, but they still walk through the world on their own. We present an extension for a video browsing tool that also follows this principle. The users still have to look for the searched segments manually, but in addition they can take advantage of further information from augmented navigation bars.

2 Augmented Navigation Bars

The VBS 2012 showed that tools like the AAU Video Browser[8] are very powerful in fulfilling certain KIS tasks. Scenes that are significantly different from other scenes in a video or scenes that are expected at certain parts of a video (e.g. in most cases weather reports are at the end of news videos) can be found fast. On the other hand, in videos that show similar content from the beginning to the end (e.g. TV shows) or videos that consist of repeating similar situations (e.g. videos where anchorpersons or some sports acts are shown again and again) scenes are hard to find only by human observation.



Fig. 1. Example of an augmented navigation bar

Therefore, we present augmented navigation bars to provide additional information to the users to make such search tasks easier. Augmented navigation bars are annotated with colored blocks, where each color represents a certain class of video content. An example is shown in Figure 1. The idea is (1) to automatically detect repeating segments within videos, (2) to cluster them in groups of similar content and (3) to annotate the navigation bars with colors indicating the location of the clusters in the video. Each cluster is visualized by a different color.

A large number of analysis methods can be used for step 1, e.g. global visual features, local visual features and audio features. In this paper, we investigate classes based on color and motion analysis. Using other descriptors may provide a more fine-grained segmentation, but this is topic for further research.

A latent indexing of the content is performed, thus no predefined classes are used. The classification of the emerging clusters is the part where the user comes into the loop. A preview panel shows the cluster centers, each one surrounded

¹ e.g. layar - <http://www.layar.com/> (2012-10-09).

² Project Glass - <https://plus.google.com/+projectglass> (2012-10-09).

with a colored border. The border indicates the color that is used for marking all segments of the corresponding cluster on the navigation bar. The preview panel is shown right to the video windows in Figure 2. It helps to make a basic discrimination of the content of a video. If the representative frames are not discriminative enough, all segments of a cluster can be loaded in a separate playlist by clicking on the cluster center. The elements of a playlist are ordered chronologically, thus scanning the items of a playlist from the beginning to the end is also an option to search for a certain video segment. A playlist can be seen at the right side of the window in Figure 2.

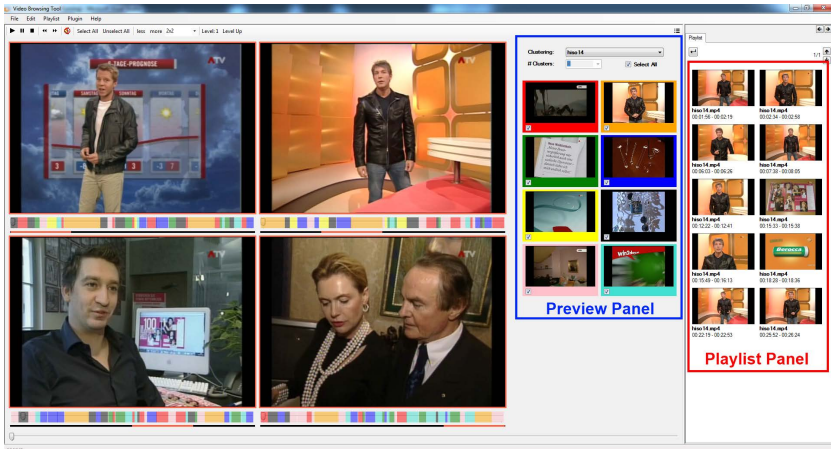


Fig. 2. The four windows at the left side can be used to browse four parts of a video in parallel. Next to them the preview panel is displayed. On the right side the playlist panel shows the segments of one cluster. Below each video window an augmented navigation bar is displayed.

The most important point regarding the presented video browsing application is that users are still interacting with videos and not only with static key-frames. Therefore, they experience videos in the same way as usual, but additionally the search space for KIS tasks can be reduced by the indication where content of different clusters is located.

3 Content Analysis, Pattern Detection and Clustering

We do not use predefined, trained classifiers for the clustering, because such methods are always dependent on the type of data used for the training. The method proposed is applicable to any video without any training in advance. Only a clustering of video segments that show similar, repeating patterns of color or motion is performed. We rely on global features, because they are sufficient to segment videos into a few basic clusters, which are easily distinguishable for users.

The color analysis is based on the color layout descriptor[17]. For the motion analysis the motion histogram introduced by Schoeffmann et al.[21] is applied. For each frame of a video the color layout and the motion histogram are extracted. Coherent video segments of similar color or respectively similar motion are detected by iterating with a sliding window of the length L over all frames f . If the difference at frame n exceeds an empirically estimated threshold T_{sim} a segment boundary is detected (d_{L2} is the Euclidean distance between two frames):

$$\frac{\sum_{i=n-L}^n d_{L2}(f_i, f_{i+1})}{L} < T_{sim} \quad (1)$$

This segment detection can be compared to shot detection, as the resulting segments often correspond to video shots. After the color and motion segments have been detected, we apply a k-means++ clustering algorithm[1] to group them into k classes. The similarity between two different video segments x and y is estimated as follows ($len(x)$ indicates the number of frames in segment x):

$$Sim(x, y) = \frac{\sum_{i=0}^{len(x)-1} \sum_{j=0}^{len(y)-1} d_{L2}(x_i, y_j)}{len(x) * len(y)} \quad (2)$$

Each class contains segments with similar color layout patterns or respectively motion layout patterns. Based on this latent classification of the video segments, the annotation of the navigation bars is performed.

4 Experimental Results

We conducted some experiments to find out how users can be supported in video browsing tasks by the two content analysis methods mentioned above. Here we discuss our findings based on the results of these experiments. We did not perform a user study so far, because we wanted to investigate to what extent the proposed approach is able to produce clusters of semantically meaningful video segments. Nevertheless, we conducted some tests on our own to investigate the potential of our proposed interface. These subjective tests showed that our tool is able to improve the search performance in KIS tasks. Based on these observations and the results of our experiments we are going to perform a representative user study in future.

We used 10 TV recordings from 5 different types of videos (2 videos of each): a society magazine, ski jumping competitions, alpine skiing competitions, soccer games and a quiz show. We did not preprocess the TV recordings before our study (e.g. we did not remove commercials or the shots of preceding or following programs), because in real use cases people usually do not preprocess their recordings before browsing them. Therefore, this study gives a good impression how our approach is applicable to day-to-day usage scenarios.

Table 1. Correlation of the clustering results

Video	Color-Based Clustering	Motion-Based Clustering	
Society Magazine 1	Anchorman Shots	0,90	Reports (High Motion) 0,80 Interviews, Anchorman (Low M.) 1,00 3 Outlier Shots 0,00 1 Outlier Shot 0,00
	Dark Shots	0,99	
	White Shots	0,82	
	Red Shots	0,82	
	Outdoor/Day Shots	1,00	
	Program Preview	1,00	
	Channel Inserts	0,71	
	3 Outlier Shots	0,00	
Society Magazine 2	Anchorman Shots	0,78	Reports (High Motion) 0,69 Interviews, Anchorman (Low M.) 1,00 19 Outlier Shots 0,00 2 Outlier Shots 0,00
	Dark Shots	0,94	
	White Shots	0,64	
	Yellow Shots	0,75	
	Skin Color Shots	0,83	
	Outdoor/Day Shots	0,60	
	Program Preview	1,00	
	5 Outlier Shots	0,00	
Ski Jumping 1	Hill Shots 1	0,97	Before/After Jump (Zoom) 0,91 Slope Shots 1,00 Close-Up Shots (Few Motion) 1,00 Start & Result List (No Motion) 1,00 Athletes slowing down 0,96 13 Outlier Shots 0,00
	Hill Shots 2	0,96	
	Sky/Snow Background	1,00	
	Spectators, Interviews, Studio Sh.	0,74	
	Blue Shots	1,00	
	Program Preview	1,00	
Ski Jumping 2	Dark Background	0,85	Close-Up, Slow Motion, Interviews 0,86 Slope Shots 0,95 Athletes After Jump (Few M.) 1,00 Start Shots, Studio Shots (No M.) 1,00 3 Outlier Shots 0,00
	Snow Background	1,00	
	Sky Background	1,00	
	Dark + Light Background	1,00	
	Program Preview	1,00	
	Channel Inserts	1,00	
Alpine Skiing 1	Snow Background	0,95	Race Shots (To The Right) 0,87 Race Shots (To The Left) 1,00 Race Shots (Downwards) 0,97 Start & Finish Shots (Zoom) 0,98 Close-Up Shots (Few Motion) 0,88 Moderator, Interviews (No M.) 1,00
	Dark Background	0,93	
	Snow + Dark Background	0,90	
	Start Shots	1,00	
	Non-Race Shots	0,86	
	Program Preview	0,83	
Alpine Skiing 2	Start Shots, Interviews	0,67	Race Shots (To The Right) 0,94 Race Shots (Downwards & Right) 0,71 Race Shots (Downwards) 0,78 Close-Up Shots (Zoom) 0,83 Overview Shots (Few Motion) 1,00 Moderator, Interviews (No M.) 0,93
	Race Shots (White)	0,97	
	Race Shots (Grey)	0,95	
	Race Shots (Yellow)	1,00	
	Race Shots (Blue)	1,00	
	Program Preview	0,90	
Soccer 1	Field Shots	0,98	Action Towards Right Side 0,94 Action Towards Left Side 1,00 Studio Shots, Moderators (Few M.) 1,00 Close-Up, Commercials (Other M.) 0,89
	Close-Up & Studio Shots	0,89	
	Interviews, Commercials	0,96	
	5 Outlier Shots	0,00	
Soccer 2	Field Shots	0,92	Action towards right side 1,00 Action towards left side 0,92 Studio Shots, Moderators (Few M.) 1,00 Close-Up, Commercials (Other M.) 0,84
	Close-Up & Studio, Commercials	1,00	
	Program Inserts	1,00	
	14 Outlier Shots	0,00	
Quiz Show 1	Close-Up (People) Shots	0,99	Question Shots 0,61 Non-Question Shots 0,84 2 Outlier Shots 0,00 1 Outlier Shot 0,00
	Overview Shots	0,46	
	Candidate Selection Shots	0,94	
	Channel Inserts	1,00	
	Commercials (Green)	1,00	
	Commercials (White)	0,71	
	Commercials (Yellow)	0,80	
	Commercials (Blue)	1,00	
Quiz Show 2	Close-Up Shots (People)	0,99	Question Shots 0,70 Non-Question Shots 0,78 35 Outlier Shots 0,00 1 Outlier Shot 0,00
	Overview Shots	0,88	
	Candidate Selection Shots	1,00	
	Commercials (Green)	1,00	
	Commercials (White)	1,00	
	Commercials (Yellow)	1,00	
	Commercials (Blue)	0,67	
	3 Outlier Shots	0,00	

We manually created a ground truth after the initial video segmentation, by assigning the detected segments to different semantic classes. The clustering started with 12 clusters for each video. In successive experiments we lowered the number of clusters for each video as long as the algorithm produced two or more clusters of the same class. At the end, the number of clusters converged to different k for different types of videos and different content features. Table 1 shows the amount of clusters each video type converged to and the correlation of the clustered items. The correlation expresses the ratio of shots in each cluster

that belong to the same group in our ground truth. Of course, in a day-to-day usage scenario it would not make sense to manually classify all segments in advance. Instead, several clustering rounds have to be performed, each resulting in a different number of clusters. At the end, the users can decide how many clusters should be displayed.

The color-based clustering is well suited for the discrimination of anchorman shots and report shots in the society magazine videos. The anchorman is always standing in the same TV studio. The background color of the studio dominates the whole shot and thus color-based clustering fits very well for finding these shots (cp. *Anchorman Shots*). Finding anchorman shots is very helpful, because these shots always occur in front of a report and thus users can quickly navigate from report to report. Furthermore, color-based clusters can be used to find other shots with a dominant color in the background or to distinguish between day or night scenes. The motion-based clustering of the society magazine can only be used for the discrimination into two classes of shots: (1) report shots, which contain motion, and (2) interview and anchorman shots, which contain no or only few motion (cp. *Interviews, Anchorman*). People watching such programs are especially interested in the celebrities shown and thus browsing from interview to interview may be an usage scenario for many users.

Ski jumping videos are characterized by repeating, very similar situations. This circumstance leads to very good clustering results for both methods. The color-based clusters are dominated by the background of the shots (snow, sky or dark background). The motion-based clustering produces groups of repeating patterns of camera zooms or pans. For example, if an athlete is running down the slope, it produces always the same motion pattern and thus users can easily navigate from one athlete to the other. An excerpt of this cluster is shown in Figure 3. Zoom motion is an indicator for close-up views of athletes, which are often shown before or after a jump. Interviews are contained in clusters with few or no motion and thus users can quickly browse all interview scenes based on that information.

For alpine skiing videos the color-based clustering leads to similar results like for ski jumping videos. Again the background of the shots dominates, but also different lighting conditions influence the clustering results. As the athletes start in front of the same significant background, i.e. a red wall in our test videos, it is also possible to group those shots where athletes start their runs. Figure 3 shows some examples of this cluster. Moreover, skiing videos contain a lot of motion. The motion-based clustering also leads to several clusters with race scenes that contain different motion directions (e.g. to the left, to the right or downwards motion). Furthermore, zooming motion indicates close-up shots of athletes, which often occur before or after a run and shots containing no motion often indicate interview scenes.

Soccer videos consist of many similar shots throughout the whole video, only a few classes can be distinguished. The color-based analysis produces groups of game scenes and close-up shots. The motion-based approach is also able to detect close-up shots. The indication of close-up shots may help users to find interesting scenes in a soccer game, because close-up views often occur if something

extraordinary happens (e.g. fouls, free kicks or goals). Furthermore, clusters of shots emerge showing motion to the right and motion to the left. All scenes where a team is running towards the left goal are contained in the same cluster. The same holds for action towards the right goal. Therefore, when people are searching for a certain offensive scene of a team the search space can be reduced with the help of the motion information.

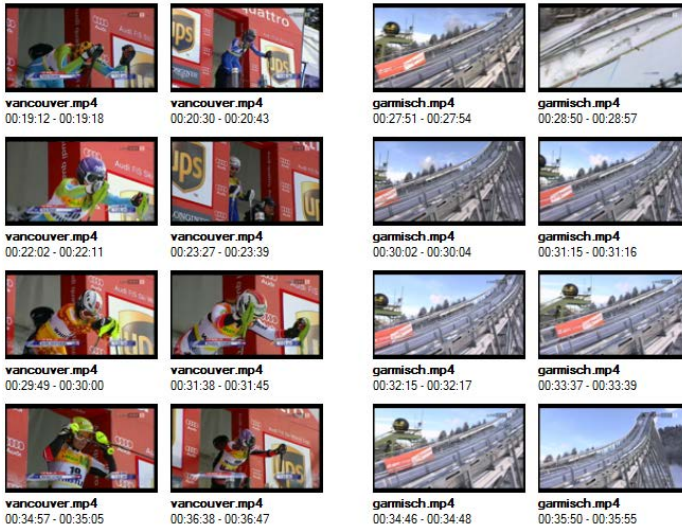


Fig. 3. Excerpts of two clusters: color-based (left) and motion-based (right)

The quiz show recordings contain a lot of commercials. Due to this, the color-based clustering produces a lot of different clusters of commercials having different dominant colors. Only three clusters emerged that contain quiz show shots. A very large one contains close-ups of the moderator and the candidates and thus all question scenes. A second one shows overview shots of the studio and a third cluster contains all scenes where a new candidate is selected. Especially, the last cluster may be helpful for users to see at a glance when a new candidate comes in. The motion-based approach is not that well suited for the quiz show videos. It seems that they contain too little motion to produce discriminative results. The only observation we made is that the motion information can be used to distinguish between two classes: (1) shots where a question is displayed to the candidate and (2) other shots. But the correlation of the clustered shots is rather low compared to the correlation of the clustering results that can be achieved for the other four video genres.

For some of the videos we received clusters with outlier shots (cp. *Outlier Shots*). These clusters contain shots that should belong to other clusters from a semantic point of view. As only a very small amount of shots belongs to these clusters, we ignore them at first.

5 Related Work

Several works have been proposed for enhancing navigation bars for video browsing tasks. Schoeffmann et al.[23] present visual navigation bars. The navigation bar is replaced by a visual representation of the video content. For example, users can choose between a dominant color or a motion representation. Users can also search for patterns in these representations, but they have to find such patterns on their own, whereas we simplify the browsing experience by reducing it to the most dominant recurring patterns in the color or motion information. Barbieri et al.[2] introduce a navigation bar based on the dominant colors of single frames, but again it is up to the user to identify patterns.

Moraveji[19] enhances the navigation bar with distinctive colors that are used to mark segments where persons, faces, vehicles, etc. are shown. In contrast to our approach they use pre-trained classifiers to detect these segments. We perform a clustering into latent classes, mark them on the navigation bar and let the users classify the content. Therefore, our approach should be applicable to more general use cases. Chen et al.[5] present the EmoPlayer, a video player that shows the emotions of the involved people on the navigation bar. The annotation has to be done manually for each video.

Chang et al.[4] introduce a filmstrip view for video browsing. At the bottom of the screen a filmstrip showing the key-frames of a video is displayed, which allows browsing and seeking in the video. The Mitsubishi Electric Research Laboratories propose key-frame-based layouts for improved fast-forward and rewind in videos[10]. Wittenburg et al.[25] generalized this visualization technique to rapid serial visual presentation (RSVP). Hauptmann et al.[13] use a RSVP-based approach for video retrieval tasks. Like our approach, they try to make maximum use of human perception skills. The results of a query are presented to the users with an RSVP-based approach. Our video browser can be used to examine several parts of a video in parallel. Therefore, users can also get a fast overview of videos by scrolling through several video parts in parallel at the same time.

Eidenberger[11] uses a self-organizing map (SOM) to group visually similar shots in hierarchically organized index trees, which can be used to browse the content. Barecke et al.[3] also use a SOM-based approach. We perform a similarity-based clustering as well, but in contrast to these two approaches the temporal order of the clustered shots is preserved for the users, which should help to better understand the context of each shot.

Several approaches have already been presented that allow browsing of video content based on events identified in the audio stream. Friedland et al. present a tool for browsing sitcoms based on detected punchlines[12]. Divakaran and Otsuka[9] propose a browsing interface for VCR devices, which enables browsing of sports videos based on cheering of the audience and excited speech of the commentator. A similar approach for racket sports is introduced by Liu et al.[18]. Investigating the audio stream also seems to be an interesting option for our approach, but so far we only concentrated on visual features.

Hürst et al.[14][15][16] also present an enhanced navigation bar, which provides different granularities for browsing the content of a video. Users can smoothly change between different granularities while moving the slider.

A tool, which combines different modalities with intelligent interaction means, is proposed by de Rooij et al.[6] and Snoek et al.[24]. They introduce the notion of video threads. Each thread represents a sequence of similar shots, where the similarity can be based on different modalities like visual similarity, similarity of textual annotations, semantic similarity, and temporal closeness. Users can choose the thread that best meets the characteristics of a specific browsing or search task. In contrast to our approach, users can browse videos only by looking at key-frames. Our tool preserves a video-player-like experience, because the users can still watch and interact with videos.

This incomplete listing of video browsing approaches mentions only papers related to our work. An extensive survey of video browsing applications is presented by Schoeffmann et al.[20].

6 Conclusion

We presented an approach for smart video browsing with augmented navigation bars. We extract small segments based on coherent color, respectively motion directions, from videos and cluster these segments based on their similarity. The navigation bars of our video browser are augmented to visualize where segments of each cluster are located in the video. Aim is to reduce the initial search space for a number of video browsing tasks, in particular for videos with recurring similar scenes [7].

The combination of well-known video interaction concepts with the results of content analysis has several advantages. Classes of similar content are suggested to the users, thus they can get a first overview of the content of a video. Compared to other key-frame-based tools the users can still watch and interact with a video as they were always used to. The augmented information only provides an additional help for them. Furthermore, the temporal order of the content is preserved, thus users always have an overview of the temporal correlation of different segments.

The evaluation showed that our approach is generally applicable to different video genres, without the need of a training phase for a specific dataset. We only investigated five video types, but our approach may also be applied to further types having repeating visual patterns. Subjective tests showed that the search performance can be improved for KIS tasks.

Based on the findings of our initial experiments and tests we are going to conduct a real user study where we want to measure the search performance in day-to-day video browsing scenarios. Furthermore, we submitted this approach to the Video Browser Showdown at MMM 2013, where we would like to measure the performance of the presented tool in KIS tasks competing with video browsing tools from other research groups.

We are also going to investigate the integration of further analysis methods, such as local feature descriptors, face detection methods and audio analysis.

Clusters of similar objects, faces and sounds may provide an additional benefit for the users. In this paper, we relied on color and motion information alone. A combination of different features may also improve the clustering results. During our experiments we had to recognize that it is a non-trivial task to combine different modalities in order to achieve better results than using one modality alone. In our case the results even got worse compared to the approaches where only one modality was used. Therefore, we leave this issue open as a topic for further research.

Acknowledgment. This work was supported by Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214 22573 33955.

References

1. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, *SODA 2007*, Philadelphia, PA, USA, pp. 1027–1035. Society for Industrial and Applied Mathematics (2007)
2. Barbieri, M., Mekenkamp, G., Ceccarelli, M., Nesvadba, J.: The color browser: a content driven linear video browsing tool. In: IEEE International Conference on Multimedia and Expo., *ICME 2001*, pp. 627–630 (2001)
3. Bärecke, T., Kijak, E., Nürnberger, A., Detyniecki, M.: VideoSOM: A SOM-Based Interface for Video Browsing. In: Sundaram, H., Naphade, M., Smith, J.R., Rui, Y. (eds.) *CIVR 2006*. LNCS, vol. 4071, pp. 506–509. Springer, Heidelberg (2006)
4. Chang, L., Yang, Y., Hua, X.-S.: Smart video player. In: IEEE International Conference on Multimedia and Expo., April 23 -26, pp. 1605–1606 (2008)
5. Chen, L., Chen, G., Xu, C., March, J., Benford, S.: EmoPlayer: A media player for video clips with affective annotations. *Interacting with Computers* 20(1), 17–28 (2008)
6. de Rooij, O., Snoek, C., Worring, M.: Query on demand video browsing. In: Proceedings of the 15th international conference on Multimedia, pp. 811–814. ACM Press, New York (2007)
7. del Fabro, M., Böszörmenyi, L.: Video scene detection based on recurring motion patterns. In: 2010 Second International Conferences on Advances in Multimedia (*MMEDIA*), pp. 113–118 (2010)
8. Del Fabro, M., Böszörmenyi, L.: AAU Video Browser: Non-Sequential Hierarchical Video Browsing without Content Analysis. In: Schoeffmann, K., Merialdo, B., Hauptmann, A.G., Ngo, C.-W., Andreopoulos, Y., Breiteneder, C. (eds.) *MMM 2012*. LNCS, vol. 7131, pp. 639–641. Springer, Heidelberg (2012)
9. Divakaran, A., Otsuka, I.: A Video-Browsing-Enhanced Personal Video Recorder. In: 14th International Conference on Image Analysis and Processing Workshops, *ICIAPW 2007*, pp. 137–142 (2007)
10. Divakaran, A., Peker, K., Radhakrishnan, R., Xiong, Z., Cabasson, R.: Video Summarization using MPEG-7 Motion Activity and Audio Descriptors. Technical Report TR-2003-34, Mitsubishi Electric Research Laboratories (May 2003)

11. Eidenberger, H.: A video browsing application based on visual MPEG-7 descriptors and self-organising maps. *Int. Journal of Fuzzy Systems* 6(3), 125–138 (2004)
12. Friedland, G., Gottlieb, L., Janin, A.: Joke-o-mat: browsing sitcoms punchline by punchline. In: *Proceedings of the Seventeen ACM International Conference on Multimedia, MM 2009*, pp. 1115–1116. ACM, New York (2009)
13. Hauptmann, A., Lin, W., Yan, R., Yang, J., Chen, M.: Extreme video retrieval: joint maximization of human and computer performance. In: *Proceedings of the 14th Annual ACM International Conference on Multimedia*, pp. 385–394. ACM Press, New York (2006)
14. Hürst, W.: Interactive audio-visual video browsing. In: *Proceedings of the 14th Annual ACM International Conference on Multimedia*, pp. 675–678. ACM, New York (2006)
15. Hürst, W., Gotz, G., Welte, M.: Interactive video browsing on mobile devices. In: *Proceedings of the 15 th International Conference on Multimedia*, vol. 25, pp. 247–256 (2007)
16. Hürst, W., Merkle, P.: One-handed mobile video browsing. In: *Proceeding of the 1st International Conference on Designing Interactive user Experiences for TV and Video, UXTV 2008*, pp. 169–178. ACM, New York (2008)
17. Kasutani, E., Yamada, A.: The mpeg-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. In: *Proceedings of the 2001 International Conference on Image Processing*, vol. 1, pp. 674–677 (2001)
18. Liu, C., Huang, Q., Jiang, S., Xing, L., Ye, Q., Gao, W.: A framework for flexible summarization of racquet sports video using multiple modalities. *Computer Vision and Image Understanding* 113(3), 415–424 (2009)
19. Moraveji, N.: Improving video browsing with an eye-tracking evaluation of feature-based color bars. In: *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries*, pp. 49–50 (2004)
20. Schoeffmann, K., Hopfgartner, F., Marques, O., Böszörményi, L., Jose, J.M.: Video browsing interfaces and applications: a review. *SPIE Reviews* 1(1) (2010)
21. Schoeffmann, K., Lux, M., Taschwer, M., Böszörményi, 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 (July 2009)*
22. Schoeffmann, K., Merialdo, B., Hauptmann, A.G., Ngo, C.-W., Andreopoulos, Y., Breiteneder, C. (eds.): *MMM 2012. LNCS*, vol. 7131. Springer, Heidelberg (2012)
23. Schoeffmann, K., Taschwer, M., Böszörményi, L.: The video explorer: a tool for navigation and searching within a single video based on fast content analysis. In: *Proceedings of the First Annual ACM SIGMM Conference on Multimedia Systems, MMSys 2010*, pp. 247–258. ACM, New York (2010)
24. Snoek, C.G.M., van de Sande, K.E.A., de Rooij, O., Huurnink, B., van Gemert, J.C., Uijlings, J.R.R., He, J., Li, X., Everts, I., Nedovi, V., van Liempt, M., van Balen, R., Yan, F., Tahir, M.A., Mikolajczyk, K., Kittler, J., de Rijke, M., Geusebroek, J.-M., Gevers, T., Worring, M., Smeulders, A.W.M., Koelma, D.C.: The MediaMill TRECVID 2008 semantic video search engine. In: *Proceedings of the 6th TRECVID Workshop, Gaithersburg, USA (November 2008)*
25. Wittenburg, K., Forlines, C., Lanning, T., Esenther, A., Harada, S., Miyachi, T.: Rapid serial visual presentation techniques for consumer digital video devices. In: *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology*, pp. 115–124. ACM, New York (2003)

Human Action Search Based on Dynamic Shape Volumes

Hong-Ming Chen¹, Wen-Huang Cheng¹, Min-Chun Hu²,
Yan-Ching Lin¹, and Yung-Huan Hsieh¹

¹ Academia Sinica

² National Cheng Kung University, Taiwan

{blacksmith,whcheng,trimy,littlelight,yhhsieh}@citi.sinica.edu.tw

Abstract. In this paper, an interactive system for human action video search is developed based on the dynamic shape volumes. The user is allowed to create a search query by freely and continuously posing any number of actions in front of the Kinect sensor. For the captured query video sequence and each data stream of the human action video database, we extracted useful shape properties on the basis of space-time volumes by exploiting the solution to the Poisson equation. Different from conventional learning-based human action recognition techniques, we apply approximate string matching (ASM) to achieve local alignment for the matching of two video sequences. The experiments demonstrate the effectiveness of our system in support of the user's search task.

1 Introduction

Traditional video search task is accomplished by giving a proper keyword query, and the performance of text-based search drastically relies on the quality of video tags/annotations. Unfortunately, many of the videos on the internet are not well annotated by the user who uploaded them, resulting in poor search results. Recently, query-by-example (QBE) has been proposed as another image search scheme [1] that analyzes the content of a query image and compares the corresponding similarity with other candidate images in various feature spaces, such as color, texture, and shape. However, for video search task, the prerequisite of an available video example is often too strong and unrealistic for practical application scenarios. Instead of using an existing example at hand, sketch based search allows users to create their own examples as queries by sketching major curves of the target image or video in their mind [2]. But using freehand drawings would be inefficient and difficult to describe the dynamic properties of the video data. To go beyond the limitations of the existing search methods, we propose a search scenario as shown in Figure 1. In our search scenario, the user can directly perform the target action in front of the camera sensor, and the recorded action video sequence is taken as the query example for search. Highly related human actions are suggested to the users after comparing the similarity of human body poses between the user's action sequence (above) and each action video sequence in the database (below). Meanwhile, since the computer vision literature [3] has

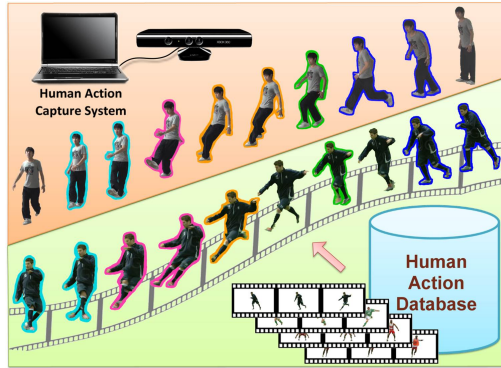


Fig. 1. The proposed human action video search scenario

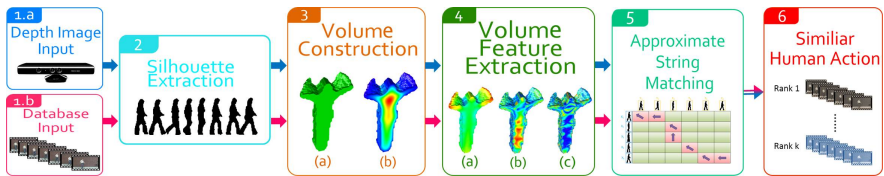


Fig. 2. The proposed system flowchart

shown that the estimation of human body pose can highly benefit from 3D depth data, we use Microsoft's Kinect and the developed API to efficiently acquire human silhouettes.

The demands for intelligent surveillance systems have advanced the action recognition techniques [4,5]. Different from action recognition systems which are designed to recognize certain predefined actions (such as walking, jumping, or falling), our system is a search based system using string matching techniques. Therefore, no action has to be predefined and arbitrary actions could be found through our system. Figure 2 shows the flowchart of the detailed processing steps in our system. First, human body silhouettes are extracted from data streams of both the depth camera and the database videos, respectively. Second, in order to fully exploit the spatial-temporal nature of human shapes for each action, consecutive silhouettes of the same person in a given length period of data stream are concatenated and thus a three dimensional space-time volume is formed. Inspired by a recent approach [6], a set of features representing the shape of the corresponding space-time volume is extracted and stored as a feature vector. The similarity between feature vectors of the query volume and the ones of the database volumes can be effectively measured using the approximate string matching techniques. Finally, highly related actions appearing in the database videos can be found and suggested to the users.

The rest of this paper is organized as follows: First, we introduce the shape-based representation in Section 2. Section 3 details the similarity measurement of two action video sequences and Section 4 demonstrates the experimental results of the proposed approach. Finally, Section 5 concludes this write-up.

2 Human Action Shape Representation

In order to accurately describe an action, we need an effective method representing this action. In the literature, most video analysis techniques treat a video sequence as a collection of still images, extract representative key frames from them, and compare their low-level features [7,8]. However, these methods relying only on key frames inherit no motion information which is of importance for action representation. Moreover, key-frame based approaches usually heavily rely on color information, but the use of color features alone is not so effective to represent an action. For example, users with different skin colors and different dressing colors can act completely the same action. In comparisons to the key-frame based approaches, a unified analysis of spatial and temporal information using a three dimensional space-time volume shows more effectiveness when describing actions. To be precise, a 3D space-time volume in (x, y, t) space is constructed by stacking consecutive spatial 2D body silhouettes extracted from each frame and thus the temporal dimension t is also taken into account. A space time volume can be formulated as:

$$V = f(s, t) = [x(s, t), y(s, t), t]. \quad (1)$$

By extracting descriptive features from the space-time volumes, we can represent an action with focus on shape and motion information. Various approaches for extracting representative features from the space-time volume have been proposed in previous works [4,9,10,11]. In [9], the notions of interest points of image are extended into spatial-temporal domain. By using second moment matrices integrated over a space-time (x, y, t) Gaussian window, a scale-invariant Harris-Laplace corner feature descriptor is derived. Analyzing these space-time corner feature points can give understanding of actions in recorded data. In [10], the authors track the contours across time dimension and generate a 3D action trajectory volume. Then they analyze the volume by using the differential geometric surface properties to identify action descriptors which capture both temporal and spatial properties. Unfortunately, since only the interest points from the surface of a volume are used for volume representation, these popular interest-point based techniques fail to represent actions in two situations. First, when an action only contains smooth motions, the surface of the corresponding volume could only have few sharp extremes and as a result, insufficient useful space-time interest points are available for feature representation. Second, the performance of these methods extremely depends on the performance of silhouette (object) segmentation from its background. However, the segmentation issue is still an open problem in the computer vision area. Occlusions, changing illumination, reflectance and shadow can easily result in the defeat of segmentation and produce

many noises. The noise upon extracted silhouettes will cause a large number of extreme values on the surface of a volume and greatly affect the surface feature representation. To avoid the above problems, in this paper, we adopt the approach proposed in [4] to extract representative features, in which the authors exploit the space-time volume and extract representative features inside the volume, rather than on the surface and hence, this approach has stronger resistance to imperfect segmentation. The shape analysis method developed for 2D data [6] is generalized to 3D volume analysis, and the solution U to the Poisson equation is utilized to obtain representative features. Considering a space-time volume V surrounded by closed surface, the Poisson equation can be formulated as follows:

$$\Delta U(x, y, t) = -1, \quad (2)$$

with $(x, y, t) \in V$, Δ denotes the Laplacian operator, and $\Delta U = U_{xx} + U_{yy} + U_{tt}$ is subject to the Dirichlet boundary condition: $\Delta(x, y, t) = 0$ at the bounding surface δV . Besides, Neumann boundary condition $U_t = 0$ is imposed to cope with the boundaries at the first frame and last frame of the volume to prevent these two frames from solution attenuation. High values of U appear in the central part of the volume whereas low values of U can be found in protrusions as shown in the Stage 3.(b) of Figure 2.

By exploiting the solution U , two kinds of useful local features indicating various local volume properties are used to represent each point (x, y, t) inside the volume. The first one is the local space-time saliency feature defined as:

$$w_{\Phi}(x, y, t) = 1 - \frac{\log(1 + \Phi(x, y, t))}{\max_{(x, y, t) \in V}(\log(1 + \Phi(x, y, t)))}, \quad (3)$$

$$\Phi = U + \frac{3}{2} \|\nabla U\|^2, \quad (4)$$

where $\nabla U = (U_x, U_y, U_t)$. In the space-time volume induced by human action, high value of Φ can be seen inside the torso whereas low value usually appears inside the fast moving limbs. Consequently, the local space-time saliency feature $w_{\Phi}(x, y, t)$ located in $[0,1]$ demonstrates its emphasis on fast moving parts.

The second type local features reveal the orientation and the aspect ratio of each point inside the volume. For each point (x, y, t) , the 3×3 Hessian matrix H of the solution U is applied to estimate these local features. The eigenvectors of matrix H reflect the local principal directions, while the eigenvalues of H imply the local curvature in the direction of the corresponding eigenvectors and are inversely proportional to the length. By comparing the absolute values of the eigenvalues, we can have the relative strength of the eigenvectors. Therefore, for every point inside a volume, we can obtain its local aspect ratio and its orientation through related eigenvalues and eigenvectors.

Assume $\lambda_1 \geq \lambda_2 \geq \lambda_3$ are the three eigenvalues of H . If the eigenvalue λ_1 is much larger than the others, the strength of the eigenvector corresponding to λ_1 should be much smaller than the others. Thus the direction of this eigenvector is the informative direction at this point and the local aspect ratio will look like

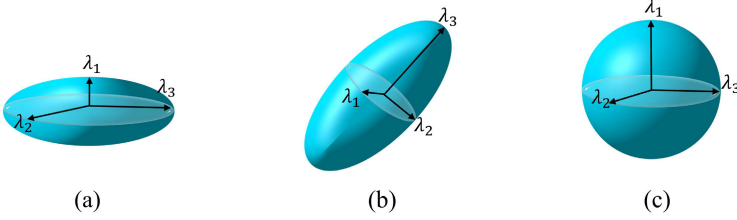


Fig. 3. Illustrations of the three types of eigenvalues ($\lambda_1, \lambda_2, \lambda_3$) compositions associated with a Hessian matrix. (a) plate structure ($\lambda_1 \gg \lambda_2 \approx \lambda_3$), (b) stick structure ($\lambda_1 \approx \lambda_2 \gg \lambda_3$), and (c) ball structure ($\lambda_1 \approx \lambda_2 \approx \lambda_3$).

a “plate”, cf. Figure 3(a). Similarly, if the eigenvalue λ_3 is apparently smaller than the others, the strength of the eigenvector corresponding to λ_3 should be larger than the others, and thus the direction of this eigenvector is the informative direction at this point. The local aspect ratio at this point will look like a “stick”, cf. Figure 3(b). Otherwise if there is no significant difference among the eigenvalues, the local aspect ratio will look like a “ball”, cf. Figure 3(c), and there is no informative direction locally.

To measure the magnitude of the three properties everywhere in a volume, “plateness” P_{pl} , “stickiness” P_{st} , and “ballness” P_{ba} are defined as:

$$P_{pl} = e^{-\alpha \frac{\lambda_2}{\lambda_1}}, \quad (5)$$

$$P_{st} = (1 - P_{pl})e^{-\alpha \frac{\lambda_3}{\lambda_2}}, \quad (6)$$

$$P_{ba} = (1 - P_{pl})(1 - e^{-\alpha \frac{\lambda_3}{\lambda_2}}), \quad (7)$$

$$P_{pl} + P_{st} + P_{ba} = 1, \quad (8)$$

where $\alpha = 3$ is suggested in [4]. Moreover, in [4], it is shown that the “ballness” property would not significantly affect the results and therefore is not adopted in our work. Consequently, we describe each point inside a volume by six local orientation feature descriptors (LOFD) indicating the informative direction at that point and the magnitude of “plateness” and “stickiness”. That is,

$$LOFD = \{w_{i,j}(x, y, t)\} = \{P_i(x, y, t) \cdot D_j(x, y, t)\}, \quad (9)$$

where $i \in \{pl, st\}$ and $D_j(x, y, t)$ means the projection magnitude of a specific informative direction j , i.e., the spatial directions x, y or the temporal direction t . In order to represent a volume by global features, we center the volume on its centroid and unify its scale in spatial domain but preserve the aspect ratio. The *weighted moment* is then used to integrate homogeneous local features, that is,

$$m_*^{p,q,r} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} w_*(x, y, t) g(x, y, t) x^p y^q z^r dx dy dt, \quad (10)$$

where $g(x, y, t)$ is the characteristic function of the volume: $g(x, y, t) = 1$ when inside of the volume, and $g(x, y, t) = 0$ otherwise. $w_*(x, y, t)$ denotes one of the seven local feature descriptors: $w_\Phi(x, y, t)$ or $w_{i,j}(x, y, t)$, and $0 \leq w_*(x, y, t) \leq 1$. p , q , and r are positive integers subject to $p + q \leq 2$ and $r \leq 2$. Therefore, a 126-dimension ($7 \times 6 \times 3$) global feature vector describing a human action in a given temporal range of the recorded sequence can be obtained.

In summary, global properties of an action space-time volume can be revealed by exploiting the solution to the Poisson equation. We adopt a set of local descriptors, i.e. a space-time saliency feature and six local orientation features derived from the Poisson’s solution to form a reliable global shape feature vector, such that we can effectively represent a human action with resistance and robustness to noises, e.g. the segmented silhouettes could be imperfect.

3 Human Action Dynamic Shape Matching

Once the representative global feature vector of a space-time volume is acquired, the next issue is how to measure the similarity between the query and each video in the database. The most intuitive approach is simply concatenating all the consecutive silhouettes and forming a single space-time volume for each recorded data stream, from either the query data or database videos. The similarity between two data streams can be then obtained by measuring the distance between the extracted global feature vectors. However, for the searching task, alignment in the temporal axis is essential before computing the similarity. For example, given a specific “pitch” action as the query, if a video sequence in the database only covers a small portion of the “pitch” action relative to the whole video, the intuitive approach will obviously fails because it inevitably takes many other unrelated actions in the same video into the space-time volume. The similarity value calculated without proper time alignment then seriously decreases even though a “pitch” action is involved in the compared video sequence. In addition, if the “pitch” action occurs several times in the compared video, we need a more sophisticated method to find all locations of related actions in the video.

In [4], the authors proposed to use a sliding window in the time domain to uniformly segment a space-time volume into so-called space-time cubes, where each cube contains eight frames (silhouettes) with an overlap of four frames between two consecutive space-time cubes. As for the query action, each of its space-time cubes is used as a sub-query to find all locations in the video database where a similar cube/action occurs. However, the eight-frame length of a single cube is too short to represent a meaningful action. An alternative solution is to set a longer cube length, but it could cause the same problem of containing multiple actions in the same cube. Inspired by the usage of approximate string matching (ASM) in the field of visual sequence matching and video copy detection [12], we adopt similar techniques to resolve our matching issue. Taking both the matching precision and the usability into account, ASM can compare two ordered-strings of symbols with error tolerance. To employ the ASM algorithm, here we also use an eight-frame-width sliding window with the same step size (four frames) to

partition a video sequence into space-time volumes as suggested in [4]. Each volume is regarded as a symbol and the global shape feature vector of the volume is used to represent the symbol. Thus, a symbol string corresponding to a recorded video sequence is generated. Assume $X = [x_1, x_2, \dots, x_m]$ and $Y = [y_1, y_2, \dots, y_n]$ denote two symbol strings with the size of m and n , respectively. The similarity between X and Y is defined as the maximum score of transforming X into Y by a sequence of operations, which can be computed using dynamic programming with the recurrence:

$$S(i, j) = \max\{0, S(i-1, j) + \delta(x_i, \epsilon), S(i, j-1) + \delta(\epsilon, x_i), S(i-1, j-1) + \delta(x_i, y_j)\} \quad (11)$$

where ϵ denotes the null symbol, $\delta(\epsilon, \mathbf{x}_i)$ is the score of inserting a symbol x_i to X , $\delta(\mathbf{x}_i, \epsilon)$ is the score of deleting a symbol x_i from X , and $(\mathbf{x}_i, \mathbf{y}_j)$ is the score of substituting x_i in X with y_j . The substitution score is calculated based on a simple linear model:

$$\delta(x_i, y_j) = c - \phi(x_i, y_j), \quad (12)$$

where c is a constant and $\phi(x_i, y_j)$ is the chi-square ground distance between two symbol values. The substitution score decision principal is that $\delta(x_i, y_j)$ would be positive if symbol x_i and symbol y_j are similar, and negative otherwise. Furthermore, negative scores are assigned to insertion and deletion operations. The local alignment can be acquired by finding the maximal score in the dynamic programming graph, and then tracing back the optimal path until encounter a zero score. Once a matched pair is found, we eliminate the corresponding volumes and do the matching procedure again with the remainder volumes. After doing string matching with the all videos in the database, we can find the top-k similar video clips for each query according to the sorted similarity scores. In addition, if either query sequence or database sequence are long, the overall matching process can be greatly accelerated by speed-up methods proposed in [12].

Applying ASM technique benefits us in several aspects. One merit of ASM is that the algorithm is quite simple and easy to implement. Second, the symbol-order information, i.e. the temporal dynamics in the shape feature vector, is preserved for both the query data and the videos in the database. In addition, a measure taking into account both the similarities and dissimilarities between feature correspondences reaches better performance than that considers similarities only. What is more important is the local alignment property of this string matching approach. By local we mean that two sequence are aligned pair-wise in the segment level rather than finding the best alignment for the entire lengths of the two sequences. This local alignment method can promise returning not merely one-to-one matching, but one-to-multiple, multiple-to-one, and multiple-to-multiple matches of segments among the two sequences under comparison. Applying the local alignment property to our system, user can freely pose any kind and any number of postures without indicating the beginning and the end of each action. Besides, similar actions with different frame rate or with different frequency are likely to obtain high similarity score based on this approach.

4 Results and Discussion

4.1 Experimental Environment and Parameter Settings

For capturing query actions and efficiently extracting sharp silhouettes of the user, we set up a Microsoft’s Kinect depth camera with a PC environment. The recent launching of the Kinect depth camera attracts programmers to develop related API libraries in different platforms. Here we adopted the OpenNI framework [13,14] for the development. We further integrated the OpenCV’s API libraries [15] to process the acquired video signals. We implemented the system using C++ and Matlab programming languages. C++ is used for processing depth data captured by Kinect camera and extracting human silhouettes from both the query data and the database videos. Matlab is applied to solve the Poisson Equation, extract volume features, and match volume sequences. All experiments are run on a desktop computer with Intel 2.83GHz Core 2 Quad CPU. All space-time volumes are brought to the same scale. We set 100 pixels as the spatial diagonal length of each volume. For the string matching process, Chi-square distance is used as the ground distance function. The insertion score, deletion score, and the constant value c in (12) are assigned empirically as -1, -1, and 1, respectively. Moreover, we observed that the results are not sensitive to the value c when it is set in the range of [0.2,1.0].

4.2 Performance of Human Action Search

We first evaluated the performance of single-action query. We collected 217 videos with different time-lengths from the UCF sports action dataset [16], the Weizmann action dataset [4], and the videos captured by our own. The numbers of videos collected from the three data sources are 51, 94, and 72, respectively. Each video simply contains one action, and all the 217 videos are taken as the video database. Fourteen users (including 9 males and 5 females) were invited to evaluate our System. Each user was asked to perform two queries. For each query, the user randomly chose two different actions from “Pitch”, “TaiChi”, and “Kick”, and mimicked to perform the chosen action in front of our system. The top-10 similar videos are then retrieved and shown to the user through an interactive interface. The user can replay the retrieved videos and check if a matched video really meets his/her query action. The overall retrieval performance was evaluated by the average precision (AP) and the mean average precision (MAP). Overall, considering the large number of collected videos in the database and the diversity of adopted action categories, our system achieves a promising performance of MAP=0.47. Moreover, Figure 4 shows the top-K precision with standard deviation among all the 28 queries. We can observe that the precision of the top-1 and top-2 retrieved videos are around 0.45, and the precision slightly decreases when K gets larger. It implies that similar actions can be successfully retrieved with high ranking by our system, i.e. roughly one out of two retrieved videos in the top-1 and top-2 results is relevant to the user’s searching actions.

As mentioned in Section 3, based on the ASM technique, the proposed system also works for the cases when a query or the database videos contain more

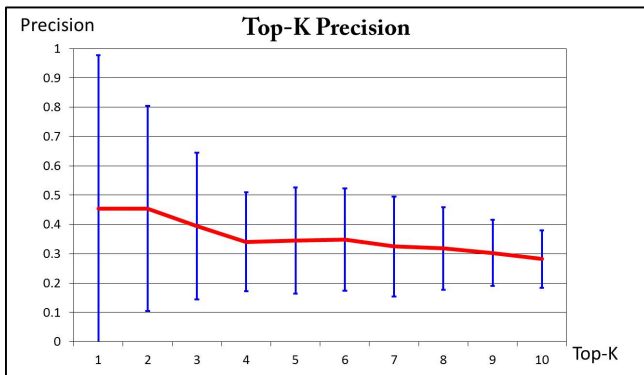


Fig. 4. The top-K precision with standard deviation among all the 28 queries. The precision of the top-1 and top-2 retrieved videos are around 0.45, and the precision slightly decreases when K gets larger.

than one action. To evaluate the system performance in such cases, we made an extended testing dataset by concatenating three action videos, which are selected from the 28 query actions performed by the 14 users, to generate a longer length video one at a time. Note that a query action is allowed to repeat within the same concatenated video. We randomly picked up 50,000 concatenated videos from all the possible combinations to form the extended testing dataset. Each of the 28 actions is then utilized as the query input to be matched with the concatenated videos containing at least one action of the same category with the query. This experiment is designed for two main purposes: 1) in this way, we can easily obtain the ground truth of the precise time-locations where similar actions occur so as to assess if the retrieved videos really match the query one, and 2) moreover, we can also evaluate if the proposed ASM technique is effective to identify all matches when multiple similar actions can be found from the same concatenated videos. Once we obtain one (or multiple) start/end points in each retrieved video, we calculate a overlap ratio of the overlapping time length between the retrieved video and the ground truth with respect to the length of the query video. A high overlap ratio means the start/end time position of the retrieved video is well-aligned with the ground truth. Figure 5 shows the histogram of the calculated overlap ratios from matched pairs, and it is obvious that more than 70% of the retrieved videos are well-aligned. Moreover, around 90% of the retrieved data have the overlap ratio larger than 50%, which means most of the retrieved clips really (or partially) contain the query action.

4.3 Application

We also conducted another experiment to reveal the practical use of our system. Two professional street dancers were invited to perform freestyle dancing actions for eight minutes continuously, and we recorded all their actions with the Kinect

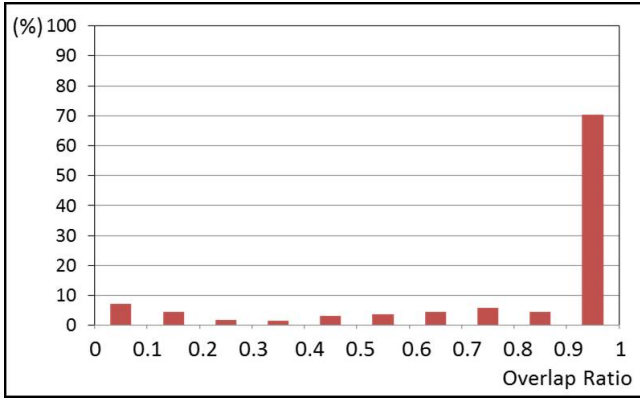


Fig. 5. The histogram of the temporal overlap ratios. See Section 4.2 for details.

camera. We let six users watch the whole video once, and asked them to manually search for two short sequences of actions randomly selected from the dancing video. In average, it will cost about 175 seconds for a user to find the both video segments. If the user wants to learn specific actions from the video, he/she has to first spend near 3 minutes finding the target actions, which is obviously quite inefficient. This problem can be easily solved by applying our system. A sequence of arbitrary actions can be directly used as query inputs, for example, if the user roughly remember some dancing actions, he can just perform them to our system for finding the corresponding video segments. It is especially useful and helpful when the actions are without precise names. As for the execution time in our current implementation (without any code optimization), averagely it takes less than 15 seconds to process a 5-second-long Kinect captured data. About 20%, 30%, 30% of the overall processing time is required for solving Poisson Equation, feature extractions, and feature moment calculations, respectively. In addition, the ASM-based matching process is near real-time. Note that the feature extraction/calculation of videos in the database can be done off-line in advance to reduce the run time spent significantly.

5 Conclusions

In this paper, we developed an interactive system for human action search in videos, in which users are allowed to create a search query by freely and continuously posing any number of actions in arbitrary orders. For each data stream of human actions, we extracted useful shape properties on the basis of space-time volumes by exploiting the solution to the Poisson equation. A set of local descriptors are derived from the Poisson's solution to form a reliable global shape feature for the action's representation and matching. The experiments demonstrate the effectiveness of our system in support of the user's search task. In

the future, we will keep our investigations along a few research directions. For example, we would like to extend our framework to hand gesture based applications, e.g. search by hand gestures or animating by hand shadows. Also, the study of free view comparisons between human actions will make our system more powerful for practical use.

References

1. Brunelli, R., Mich, O.: Image retrieval by examples. *IEEE Transactions on Multimedia* 2(3), 164–171 (2000)
2. Cao, Y., Wang, H., Wang, C., Li, Z., Zhang, L., Zhang, L.: Mindfinder: interactive sketch-based image search on millions of images. In: *Proceedings of the International Conference on Multimedia, MM 2010*, pp. 1605–1608. ACM, New York (2010)
3. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1297–1304 (2011)
4. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(12), 2247–2253 (2007), <http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html>
5. Robertson, N., Reid, I.: A general method for human activity recognition in video. *Comput. Vis. Image Underst.* 104, 232–248 (2006)
6. Gorelick, L., Galun, M., Sharon, E., Basri, R., Brandt, A.: Shape representation and classification using the poisson equation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(12), 1991–2005 (2006)
7. Ren, W., Singh, S., Singh, M., Zhu, Y.: State-of-the-art on spatio-temporal information-based video retrieval. *Pattern Recognition* 42(2), 267–282 (2009)
8. Poppe, R.: A survey on vision-based human action recognition. *Image Vision Comput.* 28, 976–990 (2010)
9. Laptev, I., Lindeberg, T.: Space-time interest points. In: *Proceedings of the Ninth IEEE International Conference on Computer Vision*, vol. 1, pp. 432–439 (2003)
10. Yilmaz, A., Shah, M.: Actions sketch: a novel action representation. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005*, vol. 1, pp. 984–989 (2005)
11. Dyana, A., Das, S.: Mst-css (multi-spectro-temporal curvature scale space), a novel spatio-temporal representation for content-based video retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* 20(8), 1080–1094 (2010)
12. Yeh, M.C., Cheng, K.T.: Fast visual retrieval using accelerated sequence matching. *IEEE Transactions on Multimedia* 13(2), 320–329 (2011)
13. OpenNI organization: *OpenNI User Guide* (2010) (last viewed January 19, 2011) 11:32
14. PrimeSense Inc.: *Prime Sensor NITE 1.3 Algorithms notes*. (2010) (last viewed January 19, 2011) 15:34
15. Bradski, G.: *The OpenCV Library* (2000) (last viewed January 19, 2011) 11:32
16. Rodriguez, M., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008*, pp. 1–8 (2008)

Video Retrieval Based on User-Specified Appearance and Application to Animation Synthesis

Makoto Okabe^{1,3}, Yuta Kawate¹, Ken Anjyo², and Rikio Onai¹

¹ The University of Electro-Communications, Tokyo, Japan

² OLM Digital, Inc. / JST CREST

³ JST PRESTO

m.o@acm.org, kawate@onailab.com, anjyo@olm.co.jp, onai@cs.uec.ac.jp

Abstract. In our research group, we investigate techniques for retrieving videos based on user-specified appearances. In this paper, we introduce two of our research activities.

First, we present a user interface for quickly and easily retrieving scenes of a desired appearance from videos. Given an input image, our system allows the user to sketch a transformation of an object inside the image, and then retrieves scenes showing this object in the user-specified transformed pose. Our method employs two steps to retrieve the target scenes. We first apply a standard image-retrieval technique based on feature matching, and find scenes in which the same object appears in a similar pose. Then we find the target scene by automatically forwarding or rewinding the video, starting from the frame selected in the previous step. When the user-specified transformation is matched, we stop forwarding or rewinding, and thus the target scene is retrieved. We demonstrate that our method successfully retrieves scenes of a racing car, a running horse, and a flying airplane with user-specified poses and motions.

Secondly, we present a method for synthesizing fluid animation from a single image, using a fluid video database. The user inputs a target painting or photograph of a fluid scene. Employing the database of fluid video examples, the core algorithm of our technique then automatically retrieves and assigns appropriate fluid videos for each part of the target image. The procedure can thus be used to handle various paintings and photographs of rivers, waterfalls, fire, and smoke, and the resulting animations demonstrate that it is more powerful and efficient than our prior work.

1 Video Retrieval by User-Specified Transformation

Because the number of accessible videos is growing larger by the day (especially on the Internet), many people are interested in ways to quickly and easily find certain scenes in these videos. Therefore, video retrieval has become an active research area for computer vision and multimedia specialists.

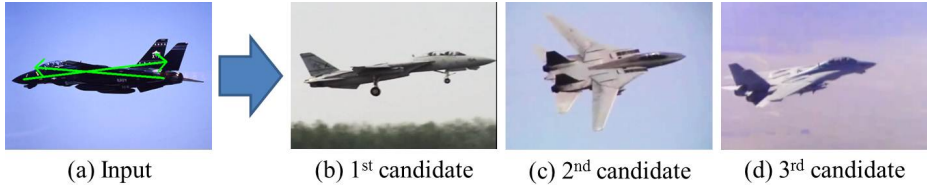


Fig. 1. (a) The current video frame and the user-drawn green arrows specifying the transformation: the tip of the fighter aircraft moves toward the right and the tail comes from the left. (b-d) The scenes retrieved from the video collection by our method, with the aircraft flying from left to right.

Most video search engines, such as YouTube¹ and gettyimages², currently support only text input in queries, and video searches are based on verbal information manually assigned to each video (e.g., the title of a video or tags). These engines do not understand queries pertaining to a desired pose or motion of a video object. If we input a text query such as “I want to watch a fighter aircraft flying from left to right,” none of the existing video search engines can retrieve scenes such as those shown in Fig. 1-b, c, or d. Furthermore, even if we were to develop a smart video search engine capable of understanding such text queries, it would be difficult or tedious for human users to precisely describe a desired pose or motion with mere words.

To address this problem, we are interested in an appearance-based user interface that enables the interactive exploration of video collections. Google Video and its extensions propose image-based image or video retrieval [1, 2] by representing each video frame as a relatively low-dimensional vector, using bag-of-features. Photo tourism and related methods allow the user to interactively explore photo and video collections by walking through a three-dimensional (3D) scene reconstructed from the collections [3–6]. These are based on precise 3D reconstruction of the scene and camera positions via the incremental structure-from-motion method. However, the technique is typically applied only to stationary objects such as buildings, is difficult to apply to moving, deformable objects, and is computationally expensive. Direct object manipulation also allows interactive navigation of scenes in a single video. [7–11]. The user navigates a video by dragging a video object and interactively editing its posture. Because these techniques are based on two-dimensional (2D) video processing rather than 3D reconstruction, their computational cost is relatively low. However, we want to perform this type of video object navigation not only in a single video, but also in video collections.

We propose an appearance-based interface that allows the user to quickly and easily specify the desired pose and motion of a video object. The user first specifies the input image by simply pausing a video or preparing some other image. Then the user specifies a transformation of an object inside the image by

¹ <http://www.youtube.com>

² <http://www.gettyimages.com>

drawing arrows, using our sketching interface (Fig. 1). After several seconds, our system displays the candidate scenes retrieved from the video database. In these scenes, the object from the input image is transformed (i.e., rotated, scaled, or translated) in the 3D world according to the user’s specifications. Nevertheless, all user input into our system is 2D, which allows the user to design a query intuitively. Our algorithm also relies only on 2D image- and video-processing technologies (i.e., does not reconstruct any 3D information), which keeps the computational cost low. We demonstrate that our method can be successfully applied to different types of video objects, including a racing car, a running horse, and a flying airplane. We also carry out a subjective evaluation of the usability of our system.

1.1 Our Approach

Let’s assume that we are now watching the scene of Fig. 2-a, where a red racing car is running from left to right. At the same time, we feel like watching another scene like Fig. 2-b, where a similar car is running toward the front. To find such a desired scene, we usually manipulate the video player many times, e.g., by pushing the forward and rewind buttons or moving the play bar: if we cannot find the desired scene in the currently watching video, we have to go to a video search engine like YouTube or gettyimages to further search for it. However, these are tedious tasks.

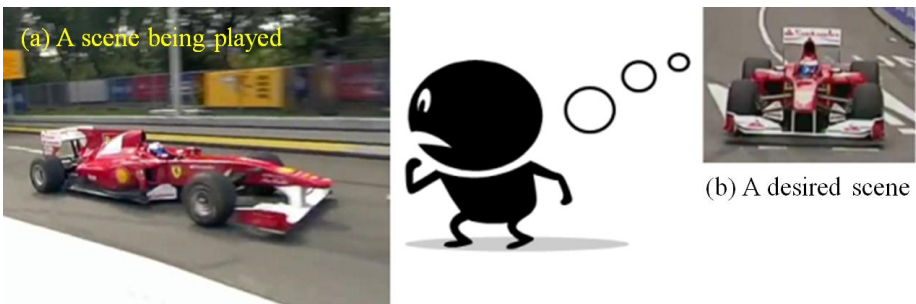


Fig. 2. Our motivation

To support this user to find the desired scene, we propose a user interaction for video retrieval. Using our sketching interface, the user specifies the transformation of the object, i.e., the red car in this case, and then the system automatically retrieves the candidates of the user-desired scene. Fig. 3-a shows the example. The user draws the two green arrows, which specify where the front and back spoilers should come in the desired scene. Fig. 3-b is the result that our system actually retrieved from the database: we overlap the same arrows over the image, which show each spoiler has come at the corresponding tip of each arrow.

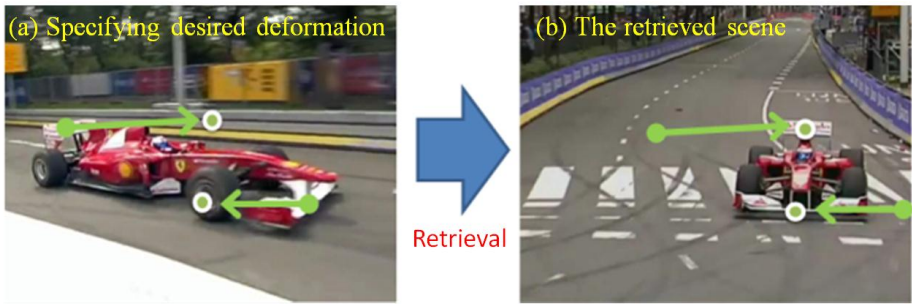


Fig. 3. The proposed user interaction

1.2 Algorithm

It is difficult for any existing image or video retrieval technique to directly retrieve the desired scene (Fig. 2-b) using the currently watching scene (Fig. 2-a) as the query. For example, using the simple method of feature matching, we extract SIFT (Scale-Invariant Feature Transform) features [12] in both the currently watching frame and every frame of the videos in the database, and compute the matching between them. Fig. 4 shows the results. In Fig. 4-left, the top and bottom frames are similar looking scenes: the red car is left-side-right but the posture of the car is almost the same. As a result, we successfully find many consistent matches between them as the many yellow lines, and the computer can retrieve the bottom frame as the result. On the other hand, there is no consistent match in the right case: the red car is the same, but since the bottom car is the rotated version of the top car. SIFT describes only 2D image feature and does not capture this 3D rotation.

However, it is interesting to note that the bottom two frames in Fig. 4 exist in the same video sequence as shown in Fig. 5. This fact lets the computer find the user-desired frame (Fig. 4-bottom right): the computer can find the frame of Fig. 5-d, and then find the desired scene by rewinding the video from the frame until Fig. 5-a is found. During the rewinding process, we track the front and back spoilers, since they are specified by the user as tails of the arrows. We stop to rewind the video, when each tracker comes near to the tip of the corresponding arrow. Then, we reach the desired scene.

In conclusion, our algorithm employs two steps (Fig. 6): we first find a frame that looks similar to the query frame, using SIFT matching, and then we automatically forward or rewind the video from that frame to find the user-desired scene. SIFT matching is efficiently performed using a kd-tree algorithm. To efficiently find the desired scene by forwarding or rewinding the video, we use the particle video algorithm [13] for motion tracking. As shown in the right part of Fig. 6, we distribute the particles throughout the image space and track their underlying motion. The red particles represent those newly added in the given

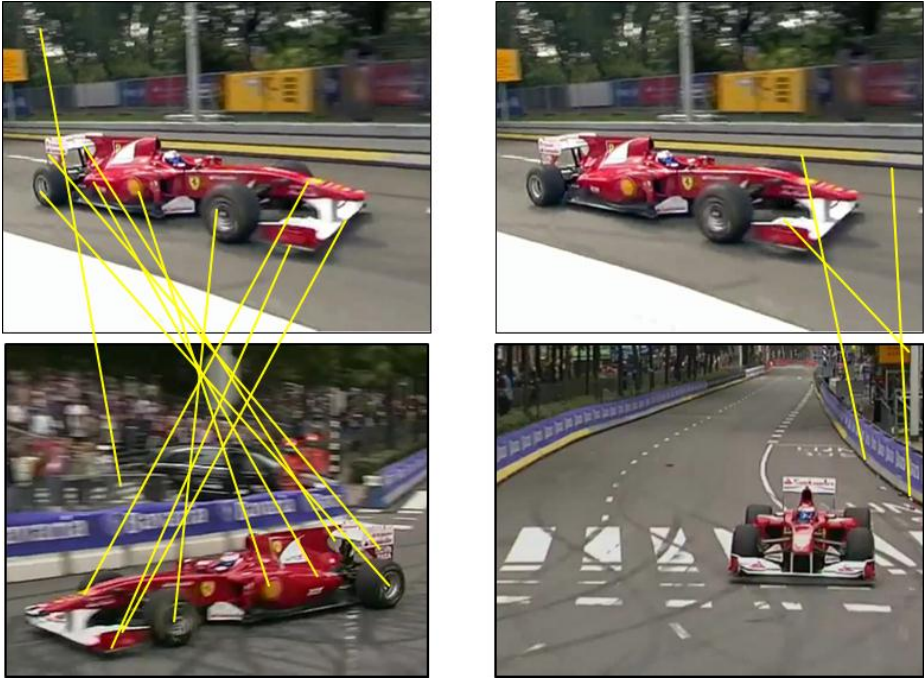


Fig. 4. Video retrieval using SIFT matching. Left: We find many consistent matches. Right: We cannot find any consistent match.

frame, while the blue particles are continuing from the previous frame. In the pre-processing stage of the database construction, all the particles and their trajectories are computed and saved. In the forwarding or rewinding process, the system selects the particles around the starting points of the user-drawn arrows in the frames found in the SIFT matching step. During the forwarding or rewinding process, the system always checks whether or not their trajectories pass through the tips of the arrows at the same time. If so, the frame corresponding to that time is output as the retrieved scene.



Fig. 5. (a) The red car is coming toward the front (Fig. 4-bottom-right), and then (d) turning to the left (Fig. 4-bottom-left)



Fig. 6. A summary of our algorithm

1.3 Results and Discussion

We tested our method on three types of video collections obtained from YouTube: videos of racing cars, running horses, and flying airplanes. All videos had a resolution of 320×240 . After downloading videos from each category, we created a video database a priori. We computed the SIFT features for every frame of each video. We also performed motion tracking using the particle video algorithm [13], and saved all particle trajectories in the database. All experiments were performed using a desktop PC with an Intel i7-860 2.8 GHz processor and 4.0 GB of memory.

Fig. 1 and Fig. 7 show the set of input images, the arrows drawn by the user, and the results of our video retrieval. Detailed statistics from our experiments are listed in Table 1. Given the user input, our system expends an average of about 4 seconds on the retrieval process.

Table 1. The statistics of our experiments. The number of video frames and the average time spent for the video retrieval are shown.

Video Type	# of Frames	Time for Retrieval (seconds)
Car	3718	4.43
Horse	4450	3.10
Aircraft	3741	4.04

The first and second rows of Fig. 7 show the retrieval of the racing car videos. In the first row, the user draws two arrows in an attempt to find scenes of a car moving from right to left. One arrow specifies that the back spoiler moves from the right, and the other specifies that the tip of the car moves toward the left. The three columns of results show the best, second best, and third best retrieved scenes. All of them show the car traveling toward the left, which matches the user's specification. In the second row, the user again draws two arrows on the front and back spoilers, to find scenes of the scaled-down car moving forward. In these results, the car has virtually the same pose in each result, but the scale varies from one result to the next. This is caused by inaccuracies in the motion tracking of the particle video algorithm; each particle often slides over the video object, especially parts moving at high speeds.

Fig. 1 shows the results for the flying airplane, and indicates one limitation of our technique. The user draws two arrows in an attempt to find scenes of a an airplane flying from left to right. In all of the retrieved results, the user’s specification has been achieved. However, in the second best candidate, the airplane is also rotated about its medial axis. Even if the user does not desire such rotation, it is often difficult to eliminate it using our system.



Fig. 7. Retrieval results

User Study. We carried out a user study to investigate the usability of our system. The subjects were nine students from the computer science department, accustomed to watching videos on the Web, but unfamiliar with our systems. We asked each of them to retrieve a desired scene using our system. We showed the car and horse images to each subject (the input images of Fig. 7), and asked the subject to envision a scene and describe it in words. Then the subject drew the arrows on the image, and the system retrieved the three best candidate scenes. The subject watched all of the retrieved scenes, and then compared them to what he/she had in mind. All subjects tried the car scene, and six of them also tried the horse scene. We asked each subject to repeat the retrieval five times, and counted the number of cases in which the subject found the desired scene.

The score for the car scene was 3.00 ± 0.70 and the score for the horse scene was 3.00 ± 1.41 . We only showed the subjects the input images of the car and the horse, and none of them had any prior knowledge of what kinds of scenes were included in the video database. As a result, subjects often specified impossible transformations of the object, and could not find the desired scene. For example, one subject wanted the car moving from the bottom to the top of the image space, but there was no aerial view of the car in our video database. The subjects

frequently commented on the difficulty of finding a number of types of scenes (e.g., a scene in which the car is running backward, showing its rear). It is actually difficult to find such a scene via a single interaction, because the rear of the car is invisible in the input image of Fig. 7, and it is impossible to specify an arrow on the rear. However, it is interesting to note that our method does allow the user to find a scene of the rear of the car by repeating the sketching and retrieval operations, rotating the car step-by-step.

2 Creating a Fluid Animation from a Single Image Based on Video Retrieval

Creating quality fluid animations is time consuming for computer graphics designers. There are two major methods. In one, physics-based simulation, it is difficult to set the appropriate physical parameters to achieve the desired appearance. The other, making a composite from a video recording of fluids, requires the time-consuming tasks of finding an appropriate video, cutting and pasting the segments accurately, and adjusting the appearance of the composite. We are interested in animating a picture of a fluid to quickly and easily achieve the desired appearance.

A previous study successfully designed fluid animations from pictures, but it was limited to synthesizing relatively calm fluid motions such as water surfaces [14]; our study focuses on synthesizing more dynamic motions such as water splashes. Another method allows users to specify a video example and then transfers its fluid features to the target image [15]; however, only a single video example is used, which limits the available variation in fluid features.

To address these problems, we developed a data-driven method for creating a fluid animation from a picture (Fig. 8). The user inputs a target image (Fig. 8-b) with a few hints about fluid motion (i.e., flow direction and speed) such as sketches of flow direction, shown as orange arrows. The user also specifies an alpha matte that extracts the fluid region of interest (Fig. 8-c). We constructed a video database that includes hundreds of video examples of fluids (Fig. 8-a) and helps the user synthesize better quality animation with less effort than in previous methods (Fig. 8-d). The technical detail is described in our paper [16].

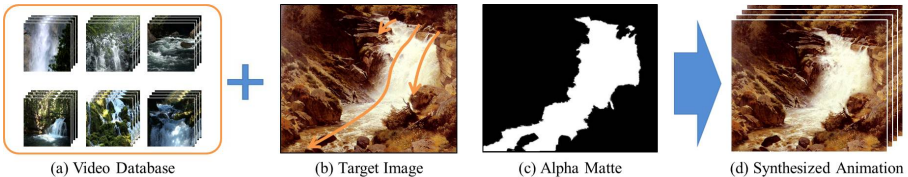


Fig. 8. Creation of a fluid animation from a picture

2.1 Our Method

Our system consists of three components: 1) construction of a video database of fluids (Fig. 9-a), where each video example is cut into small pieces; 2) a best-match search for an appropriate video example piece and assignment of this to part of the target image; and 3) synthesis of the final animation through seamless integration of the assigned pieces and adjustment of the overall appearance.

The offline process of database construction begins with gathering original video examples of fluids (Fig. 9-b). To increase the number of examples, we cut each video example into small pieces (Fig. 9-c). For each video example piece, we then compute the average image by averaging the frames (Fig. 9-e) to obtain representative information about its appearance. We also calculate the differences between the average image and frames (Fig. 9-f) that have no significant color properties but that capture high-frequency fluid features. Finally, from the averaged images in the database, we construct a bag-of-features codebook and describe each average image using a histogram of visual words (Fig. 9-d).

Hence, we cut a target image (Fig. 9-g) into small pieces using the same process used for database construction (Fig. 9-i). Next, we compute the histogram of visual words for each piece (Fig. 9-h), perform a best-match search between histograms of visual words (see Figs. 9-d and h), and assign video example pieces that are similar to the target-image piece. When a user-specified motion field is given, it is used as a constraint for solving the assignment problem. Based on the assignment results, differences (Fig. 9-f) are copied onto the corresponding target-image piece (Fig. 9-j). Finally, all assigned differences are integrated seamlessly, and the animation is synthesized by adjusting the appearance (Fig. 9-k).

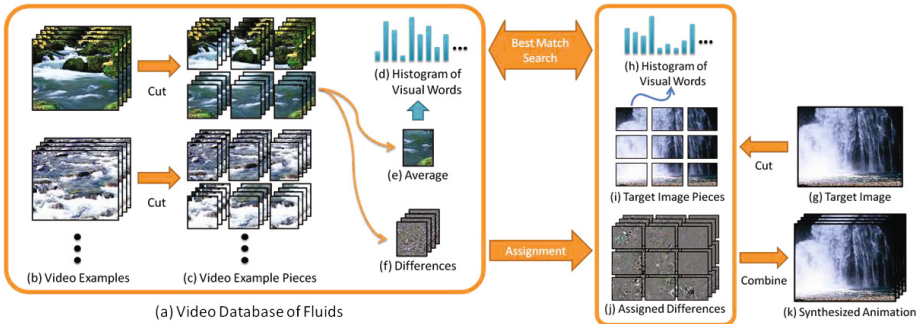


Fig. 9. System overview

2.2 Results

We constructed an independent database for water, fire, and smoke scenes. This involved gathering 151, 96, and 89 video examples for the water, fire, and smoke databases, respectively, from which we obtained 246, 227, and 195 thousand video example pieces. We synthesized the fluid animation for each target image,

as shown in the supplementary video. We designed an alpha matte and specified an orientation map for each target image. We specified a speed map only for the waterfall painting, the fire painting, and the smoke painting.

A side-by-side comparison shows that our method was better at reproducing the fluid features of original video examples than the previous method. In particular, the dynamic water splashes of the video example were not well reproduced in the previous method; it looks as if irrelevant synthesized noises flow along a static motion field. On the other hand, our method successfully reproduces the fluid features of assigned video example pieces. Our method also made it easier for users to create a fluid animation. For example, using the previous method, the waterfall painting had to be divided into the waterfall and river parts; in contrast, our method could process the whole image at once.

We also performed a user study in which 16 participants ranked the visual quality of each animation. All of the water scenes were given high scores, but fire and smoke scenes scored lower. The smoke in the train scene was difficult to animate because the motion of smoke became chaotic due to a failure in video assignment, and the lower smoke in the scene had visible artifacts of noise caused by video compression that were hidden in the original videos.

3 Conclusion

We have developed the appearance-based user interfaces for video retrieval and the application of the video retrieval technique to animation synthesis. In our approaches, we start with a single image as the input, and then introduce the additional user's suggestions. In the first study, it was the user-specified deformation. In the second study, it was the user-specified motion field of the fluid flow. In both studies, we relied on the sketch-based user interface, i.e., drawing the arrows. We adopted it because we thought it was intuitive for the user. We have demonstrated that the combination of content-based image and video retrieval technique with a few additional suggestions enabled us to propose a novel interaction for video retrieval and animation synthesis.

References

1. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: ICCV, pp. 1470–1477 (2003)
2. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.* 40(2), 5:1–5:60 (2008)
3. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. In: *ACM SIGGRAPH 2006 Papers*, pp. 835–846 (2006)
4. Agarwal, S., Snavely, N., Simon, I., Seitz, S., Szeliski, R.: Building rome in a day. In: *ICCV 2009*, pp. 72–79 (2009)
5. Frahm, J.M., Pollefeys, M., Lazebnik, S., Zach, C., Gallup, D., Clipp, B., Raguram, R., Wu, C., Johnson, T.: Fast robust large-scale mapping from video and internet photo collections. *ISPRS Journal of Photogrammetry and Remote Sensing* 65(6), 538–549 (2010)

6. Tompkin, J., Kim, K., Kautz, J., Theobalt, C.: Videoscapes: Exploring sparse, unstructured video collections. *ACM Transactions on Graphics (Proc. of SIGGRAPH)* (2012)
7. Kimber, D., Dunnigan, T., Girgensohn, A., Shipman, F., Turner, T., Yang, T.: Trailblazing: Video playback control by direct object manipulation. In: *IEEE International Conference on Multimedia and Expo.*, pp. 1015–1018 (2007)
8. 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: *Proc. of ACM Multimedia*, pp. 423–432 (2007)
9. Dragicevic, P., Ramos, G., Bibliowicz, J., Nowrouzezahrai, D., Balakrishnan, R., Singh, K.: Video browsing by direct manipulation. In: *Proc. of CHI 2008*, pp. 237–246 (2008)
10. Goldman, D.B., Gonterman, C., Curless, B., Salesin, D., Seitz, S.M.: Video object annotation, navigation, and composition. In: *Proc. UIST 2008*, pp. 3–12 (2008)
11. Karrer, T., Weiss, M., Lee, E., Borchers, J.: Dragon: a direct manipulation interface for frame-accurate in-scene video navigation. In: *Proc. of CHI 2008*, pp. 247–250 (2008)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision* 60, 91–110 (2004)
13. Sand, P., Teller, S.: Particle video: Long-range motion estimation using point trajectories. In: *Proc. of CVPR 2006*, pp. 2195–2202 (2006)
14. Chuang, Y.Y., Goldman, D.B., Zheng, K.C., Curless, B., Salesin, D.H., Szeliski, R.: Animating pictures with stochastic motion textures. In: *Proc. SIGGRAPH 2005*, pp. 853–860 (2005)
15. Okabe, M., Anjyo, K., Igarashi, T., Seidel, H.P.: Animating pictures of fluid using video examples. *Computer Graphics Forum (Proc. EUROGRAPHICS)* 28(2), 677–686 (2009)
16. Okabe, M., Anjyo, K., Onai, R.: Creating fluid animation from a single image using video database. *Comput. Graph. Forum* 30(7), 1973–1982 (2011)

Landmark History Visualization

Weiqing Min^{1,2}, Bing-Kun Bao^{1,2}, and Changsheng Xu^{1,2}

¹ National Lab of Pattern Recognition, Institute of Automation, CAS,
Beijing 100190, China

² China-Singapore Institute of Digital Media, Singapore, 139951, Singapore
{wqmin,csxu}@nlpr.ia.ac.cn, bingkunbao@gmail.com

Abstract. Landmark image mining and detection have been studied for many years, however, most of the existing work focuses on their spatial attributes, while largely ignoring the temporal information in specific ones, which are taken during historical moments. This kind of images are more valuable than the normal ones as they not only contain more comprehensive information to illustrate the landmarks in different moments, but also are useful in many real world applications, such as tour recommendation and history education. In this paper, we present a novel framework named Landmark History Visualization (LHV) to mine relevant and diverse images for each landmark's historic moments. There are two steps in LHV. The first one is to extract the event list of each landmark from Wikipedia. The event keywords are extracted, and some of them are automatically labeled as 3W (What, Who, When). In the second step, images searched by the landmark name are firstly collected from Flickr and Google images. Secondly, we employ manifold ranking with detected 3W to retrieve the relevant images, and lastly, an outlier detection and diversification based re-ranking approach is introduced to provide users with various returned images. We implemented our approach on 6 landmarks and the results demonstrate the effectiveness of LHV.

Keywords: landmark history, event, 3W, manifold ranking.

1 Introduction

There is a large amount of landmark images available on photo-sharing websites such as Flickr and Picasa, many of which are taken during some historical moments. Unlike the normal ones, images which are related to historical moments of landmarks are more informative and valuable for lots of real-world applications such as tour recommendation and history education. Therefore, visualizing landmark's history through these images is becoming demanding.

Extracting the event list and seeking relevant images are naturally two key steps in landmark history visualization. For event list extraction, considering that Wikipedia provides the description of each landmark including its history, it can be regarded as a good source for us. For seeking relevant images, as one of most popular image sharing website, Flickr hosts lots of landmark images taken

during special moments. However, since Flickr was just opened 8 years ago, there are plenty of images taken in the recent years but lack of those taken before 2004. Google images, which allows users to search images from web, provides us a complementary source to Flickr in our work. Thus, we extract the event list from each landmark’s Wikipedia page, and seek relevant images from Flickr and Google images. Two challenges exist in our work:

- To extract comprehensive and well-organized event list. It is not challenging to list a small number of events. However, what we need is a comprehensive and well-organized event list of one landmark, which is not easy to achieve.
- To bridge the semantic gaps over different cross-media sources. It is hard to explore the cross-media correlations between landmark events and images. Also, the event-images synchronization is difficult.

This paper proposes a new framework named Landmark History Visualization (LHV), which explores relevant and diverse images for events of each landmark. As an example, Fig. 1 shows the result of “Big Ben” from our proposed LHV. There are two steps in our work. The first one is event list extraction, and the second one is event-based image retrieval and re-ranking. In event list extraction, we extract sentences with date (SwD) from articles of Wikipedia. The sentences, which follow closely each SwD, are combined with the corresponding SwD to constitute a semantically complete event. Then event keywords are extracted and some of them are labeled as 3W elements (When, Who, What) for further retrieval of images. For event-based image retrieval and re-ranking, the goal is to find relevant and diverse images for each event. Firstly, we crawl images by searching with the landmark name both from Flickr and Google image. Secondly, we adopt manifold ranking with 3W elements (When, Who, What) to retrieve relevant images for each event. Thirdly, we introduce an outlier detection and diversity based re-ranking approach to enhance the diversity of the returns. The whole procedure is shown in Fig. 2.

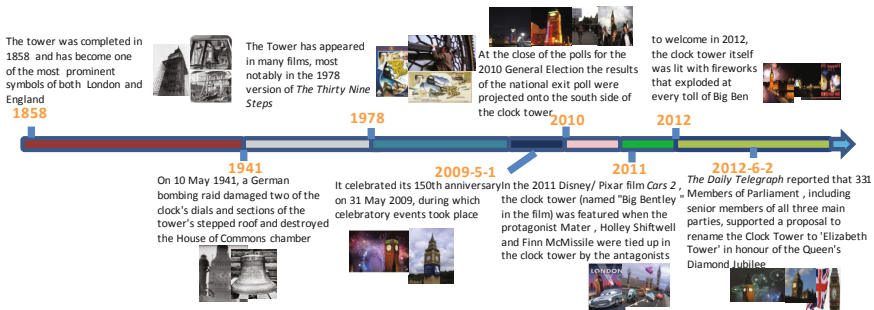


Fig. 1. Visualization of Big Ben History using our LHV

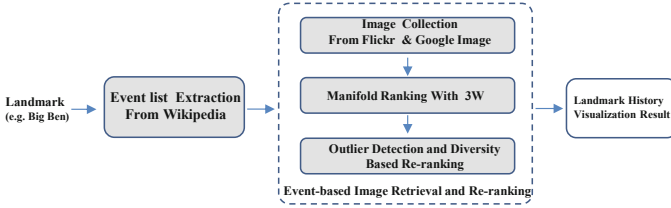


Fig. 2. The framework of LHV

The main contributions of this paper can be summarized as follows:

- Different from other work on landmark image mining and detection, our work focuses on the temporal attribution of them to explore the respective images taken in historical moments of each landmark, which are useful in many real world applications.
- We present a new framework named Landmark History Visualization (LHV), which can automatically obtain the desired event list from the landmark’s Wikipedia page and the desired images from Flickr and Google images.
- The framework can be easily extended to history visualization of other objects, such as celebrities and products.

The rest of this paper is organized as follows. Section 2 and Section 3 introduce event list extraction and event-based image retrieval and re-ranking respectively. The experiment results are reported in Section 4. We conclude the paper with future work in Section 5.

2 Event List Extraction

For each landmark, the event list is extracted from articles in Wikipedia. Firstly, similar to [5] [3] [2], we select the sentences containing temporal information since most events start from a sentence with temporal information. Temporal expressions can be grouped into four types: date, time, duration and set according to TimeML [11]. In order to extract fine-grained events, we only consider the date type, which refers to a specific point in time with a different granularity, such as “January 28, 2011”, “January, 2011”. Secondly, for each Sentence with Date (SwD), some sentences following closely their SwD may be more relevant to their corresponding SwD and thus should be considered. In particular, we employ the following rules for these sentences:

- Many events have some important named entities, such as the person and organization name. Therefore, if two sentences share the same named entities, the two sentences probably refer to the same event.

- Conjunctions and some commonly used phrases for connecting two sentences, such as “so”, “on this occasion”, reflect the author’s intention of a semantic bridge between adjacent sentences, hence these adjacent sentences may share the same significance.

In our paper, the number of considered sentences following with each SwD is set as 2. If there are more than one shared named entity, conjunction or phrase in these three sentences (SwD together with the following two sentences), all of them will be considered as the candidate sentences to event extraction, otherwise, we will only choose SwD. As defined in [1], an event generally contains some elements: Where, When, Who, What. The landmark itself is a place and represents the “Where” information, hence we extract event keywords and label remaining three elements When, Who, What (3W) for the retrieval of images in the next step.

3 Event-Based Image Retrieval and Re-ranking

After getting the event list of each landmark, we then explore relevant and diverse images for each event in the following steps: image collection, manifold ranking with 3W, outlier detection and diversification based re-ranking.

3.1 Image Collection

We crawled landmark images from Flickr by searching the names of landmarks. Meanwhile, images from Google images are also crawled for complimenting images so that these images can cover more historical moments. In addition, tags, the title and upload time of images from Flickr are downloaded. Similarly, we extracted the surrounding text, title in the page and the temporal information such as publish time of images from Google images.

3.2 Manifold Ranking with 3W

For each landmark, given the event list $E = \{e_1, e_2, \dots, e_{|E|}\}$ and images $I = \{x_1, x_2, \dots, x_{|N|}\}$, where $|E|$ and $|N|$ are the number of events and images respectively, the fundamental task is to find relevant images for each event. In fact, we can consider this task as the image retrieval problem, where each event can be the query to retrieve event-relevant images. Manifold ranking [15] [4] is employed for its effectiveness in image retrieval. It is determined by two factors: original ranking scores $\bar{\mathbf{F}}$ and transition matrix \mathbf{W} .

For $\bar{\mathbf{F}}$, the element $r(x_j)$ is generally the similarity between image’s metadata and the textual query, which is calculated by one vector modal. Different from it, we take multiple elements of an event into account and adopt a multidimensional vector modal for the similarity. Specifically, we consider 3W: When, Who and What for each landmark:

- When. Since each event happened at a special time, the same events should have the temporal proximity. Therefore, we consider the temporal similarity between time from query and image’s metadata, which is denoted as $D(e_i, x_j)$.
- Who. Some named entities such as the person name and organization name are important elements for an event. In our work, we calculate the similarity based on these named entities, which is referred to as $SW(e_i, x_j)$.
- What. Event-relevant words, such as verbs “fight”, are more important and even sometimes represent event itself. The similarity based on these event-relevant words is denoted as $SE(e_i, x_j)$.

Besides the words relevant to “who”, “whom” and “what” information, the similarity on remaining keywords is denoted as $SO(e_i, x_j)$. We detail $D(e_i, x_j)$, $SW(e_i, x_j)$, $SE(e_i, x_j)$ and $SO(e_i, x_j)$ in the following.

For temporal similarity, after the temporal information is normalized to the uniform format “yyyy-mm-dd”, we adopt the method of [10] to calculate the temporal similarity $D(x_i, x_j)$ as:

$$D(e_i, x_j) = \frac{\alpha_{poss} + 1}{\alpha_t + 1} \quad (1)$$

where α_t denotes the number of mapping steps that are needed to achieve equality for e_i, x_j . For example, consider two dates “1999-03-19” and “1999”, $\alpha_t(\text{“1999-03-19”}) = \text{“1999-03”}$, $\alpha_t(\alpha_t(\text{“1999-03-19”})) = \text{“1999”}$, here $\alpha_t = 2$. α_{poss} denotes the number of possible mapping steps for x_i, x_j after both time information have been mapped to be of equal granularity.

For named entity based similarity, we employ the bag of word model. Specifically, we collect all the named entities from keywords of E and tags of I to build the dictionary. The named entities in event e_i and image x_j are converted into a feature vector $\mathbf{z}_i, \mathbf{z}_j$ by traditional *tf-idf* weighing method respectively. The normalized linear kernel is used for the named entity based similarity between the event e_i and image x_j as

$$SN(e_i, x_j) = \frac{\mathbf{z}_i^T \mathbf{z}_j}{\sqrt{\mathbf{z}_i^T \mathbf{z}_i} \sqrt{\mathbf{z}_j^T \mathbf{z}_j}} \quad (2)$$

Similarly, $SW(e_i, x_j)$ and $SO(e_i, x_j)$ are both calculated through the normalized linear kernel. The total similarity between event e_i and image x_j is the weighed sum of the aforementioned similarity:

$$S(e_i, x_j) = a * SN(e_i, x_j) + b * SW(e_i, x_j) + c * D(e_i, x_j) + d * SO(e_i, x_j) \quad (3)$$

where a, b, c, d are the weights and $a + b + c + d = 1$. $S(e_i, x_j)$ is considered as the initial score of image x_j . That is $r(x_j) = S(e_i, x_j)$

The element $W(x_i, x_j)$ of \mathbf{W} is the fusion of multi-modal similarity. In addition to $S(x_i, x_j)$ from Eqn. (3), we also integrate the visual similarity $V(x_i, x_j)$

between image x_i and image x_j , which is directly computed based on Gaussian Kernel function with a radius parameter σ .

$$V(x_i, x_j) = \exp\left(\frac{\|x_i - x_j\|}{\sigma^2}\right) \quad (4)$$

Then the element $W(x_i, x_j)$ is calculated as:

$$W(x_i, x_j) = \psi_1 * V(x_i, x_j) + \psi_2 * S(x_i, x_j) \quad (5)$$

where ψ_1 and ψ_2 are weights and $\psi_1 + \psi_2 = 1$.

Finally, the relevance score \mathbf{r}^* can be solved by the following framework of manifold ranking:

$$\mathbf{r}^* = \min_{\mathbf{r}} (\mathbf{r}^T (\mathbf{I} - \mathbf{D}^{-1/2} \mathbf{W} \mathbf{D}^{1/2}) \mathbf{r} + \lambda \|\mathbf{r} - \bar{\mathbf{r}}\|) \quad (6)$$

where \mathbf{D} is a diagonal matrix and its element $d_{ii} = \sum_{j=1}^N w_{ij}$. $0 < \lambda < 1$ is the bias parameter. $\mathbf{r} = [r(x_1), r(x_2), \dots, r(x_N)]$ represents relevance scores of all images in I , whose element $r(x_i)$ denotes the relevance score of the image x_i .

3.3 Outlier Detection and Diversification Based Re-ranking

Up to now, we only consider the relevance of images for an event. In fact, diversity of images [14] [13] are very important and should also be considered. Therefore, we take diversification into account and propose an outlier detection and diversification based re-ranking approach for diverse images of each event.

Similar to [6], the formulation of outlier detection is as follows:

$$\begin{aligned} & \min_{\mathbf{c}} (\|\mathbf{s} - \mathbf{V}\mathbf{c}\|_1^2 + \alpha \|\Phi\mathbf{c}\|_1^2) \\ & s.t. \mathbf{c} \in \{0, 1\}^M \end{aligned} \quad (7)$$

where \mathbf{V} is visual similarity matrix. $\mathbf{s} = \mathbf{V}\mathbf{e}$ is a vector, whose element denotes the total similarity of each image to all other images. Each element of \mathbf{e} is 1. \mathbf{c} , a binary vector, denotes the outlier identification result. $c_i = 0$ means that x_i is an outlier while $c_i = 1$ means that x_i is a relevant image. $\Phi \in R^{N \times N}$ is a diagonal matrix, and the diagonal element is the weight for the corresponding element in \mathbf{c} . α is a trade-off parameter. The first term is to minimize the reconstruction error between \mathbf{s} and $\mathbf{V}\mathbf{c}$ while the second term is the sparsity constraints on \mathbf{c} to control the number of irrelevant images to remove.

After detecting irrelevant images using outlier detection, we remove the corresponding elements from \mathbf{r}^* to get the new ranked list $\mathbf{r}_{\text{new}}^* = [r_{x_1}, r_{x_2}, \dots, r_{x_M}]$, where M is the number of images after removing irrelevant images. Then we incorporate Average Diverse Precision (ADP) [14] into our diversification based re-ranking. Derived from Average Precision (AP), ADP can integrate diversity in addition to relevance. It is defined as

$$ADP(\mathbf{r}, I) = \frac{1}{R} \sum_{j=1}^M y(x_j) Div(x_j) \left(\frac{\sum_{k=1}^j y(x_k) Div(x_k)}{j} \right) \quad (8)$$

where $Div(x_j)$ indicates the semantic diversification score of image x_j . $y(x_j)$ is a binary vector, which indicates the relevance score of image x_j , i.e., $y(x_j) = 1$, if x_j is relevant and 0, otherwise. R is the number of true relevant images in the set I .

The i th image is decided as follows derived from Equ. (8):

$$\tau(i) = \arg \max_{x \in I - S_i} \frac{r(x)}{i} Div(x)(C + Div(x)) \quad (9)$$

where $\tau(i)$ is the image at the position of rank i and

$$S_i = \{\tau(1), \tau(2), \dots, \tau(i-1)\} \quad (10)$$

$$C = \sum_{k=1}^{i-1} r(\tau(k)) Div(\tau(k)) \quad (11)$$

Equ. (9) is determined by two key factors: relevance score $r(x_j)$ and diversification score $Div(x_j)$. $r(x_j)$ is directly from $\mathbf{r}_{\text{new}}^*$. $Div(x_j)$ can be calculated as

$$Div(x_j) = \min_{1 \leq i \leq j} (1 - s(x_i, x_j)) \quad (12)$$

where $s(x_i, x_j)$ is the semantic similarity, calculated by Eqn. (3).

Finally, we iterate Eqn. (9) to rerank images, and select the top ranked images.

4 Experiment

We select 6 landmarks shown in Table 1. For each landmark, the dataset contains its event list and images. In order to extract the event list, we use NLP tools OpenNLP [8] to split the text from Wikipedia into sentences, then employ HeidelTime [9] to detect date and extract these SwDs. In addition, 2 sentences following closely their SwDs are also extracted. All the named entities, conjunctions and relevant phrases from these sentences and their SwDs are extracted. Based on the rules in Section 2, if at least 1 of them is shared between the sentence and its SwD, this sentence is combined with SwD. Finally, the combined sentences form an event. For images, we crawl them and corresponding metadata from Flickr and Google images using the corresponding API respectively.

It is noted that not every event in the extracted event list has corresponding images in the dataset. Therefore, We firstly filter the event list. It is performed as follows: using a year as the granularity, if there are no corresponding images taken in this year in image set for an event, we remove events from the origin event list. The final statistics of collected dataset are shown in Table 1. For all extracted text information from articles and metadata, person names, organization names and the verbs are all extracted to construct different vocabularies for calculating semantic similarity in Section 3.2 respectively.

Table 1. The statistics of dataset

Landmark	#event	#image
Big Ben	11	56169
Eiffel Tower	12	61096
Statue of Liberty	11	62256
Sydney Opera House	10	25247
Tokyo Tower	11	15018
Lincoln Memorial	13	28873

As an example, Table 2 shows some extracted events for Eiffel Tower. For event #2, since the two sentences share the same person name “Hackett”, they are combined as an event. Meanwhile, the two sentences in event #3 are also combined since there is “On this occasion” between two sentences.

Table 2. Illustration of Some Extracted Events

#1	19 October 1901 Alberto Santos-Dumont in his Dirigible No.6 won a 10,000-franc prize offered by Henri Deutsch de la Meurthe for the first person to make a flight from St Cloud to the Eiffel tower and back in less than half an hour.
#2	1987 A.J. Hackett made one of his first bungee jumps from the top of the Eiffel Tower, using a special cord he had helped develop. Hackett was arrested by the Paris police upon reaching the ground
#3	New Year's Eve 1999 The Eiffel Tower played host to Paris's Millennium Celebration. On this occasion, flashing lights and four high-power searchlights were installed on the tower, and fireworks were set off all over it
#4	In January 2007, the multi-Michelin star chef Alain Ducasse was Brought in to run Jules Verne

For every image, we extract 809-D visual features [16], including 81-D color moment, 37-D edge histogram, 120-D wavelet texture feature, 59-D LBP feature [7] and 512-D GIST feature [12].

For parameter settings, in order to highlight semantic elements of the event, for a, b, c, d in Equ. (3), we set the higher weight for SN, SW, D . Specifically, the parameters a, b, c, d are fixed at 0.30, 0.30, 0.30 and 0.10 respectively. the parameter σ in Equ. (4) is set as 0.14. The γ in Eqn.6 is empirically set as 0.75. In Equ. (5), since initial text search has already used the text information as the relevance score in manifold ranking, additional performance improvement will come from the visual aspect. Therefore, we set higher weights ψ_1 as 0.85, ψ_2 is 0.15. α in Equ. (7) is fixed at 120, which is consistent with [6].

To demonstrate the performance of the proposed LHV scheme, we conduct objective evaluation and user studies for LHV.

4.1 Objective Evaluation

Since LHV is achieved mainly by event-based image retrieval, we use IR metrics to evaluate the performance. The metric ADP [14] calculated by Equ. (8) is

utilized. This is because it can take relevance and diversification of images into account simultaneously by extending conventional average precision (AP). In our experiment, we adopt Mean ADP@10 (MADP@10) for calculating the mean value of the ADP of all event queries for each landmark.

We compare our approach with the following four baselines to demonstrate its effectiveness. The baselines are as follows:

- text-based image retrieval (TIR). Here, the general one vector modal is adopted.
- multidimensional vector modal based image retrieval (MVMIR).
- multidimensional vector modal + manifold ranking (MVMIR-MR).
- multidimensional vector modal + manifold ranking + outlier detection (MVMIR-MR-OD).

Also, for every event query, we asked for 5 participants to annotate the top 10 retrieved images with “event-relevant” and “event-irrelevant” for other baselines since the retrieval result from our method has been ranked by ADP. If more than 3 participants thought the image is relevant to an event, the value of rank is set to 1 and 0, otherwise. Fig. 3 shows results of each landmark for MADP@10. We can see that MVMIR achieves much better performance than TIR. This is because the semantic elements of the event, such as temporal information, named entities, can enhance the discriminative power. MVMIR-MR performs better than MVMIR, since manifold ranking fuses visual information. MVMIR-MR-OD further improves the performance in that outlier detection can remove some irrelevant images. Among all other baselines, our method incorporating both multiple semantic elements of the event and diversification of images turns out to achieve the highest performance.

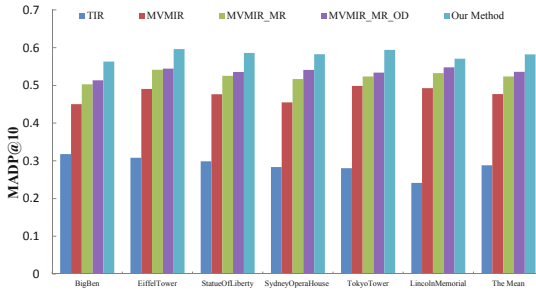


Fig. 3. Performance comparison on event-based image retrieval

4.2 User Studies on LHV

LHV provides users top five images for every event on each landmark. We also utilize the aforementioned four methods to compare with ours. There are 20

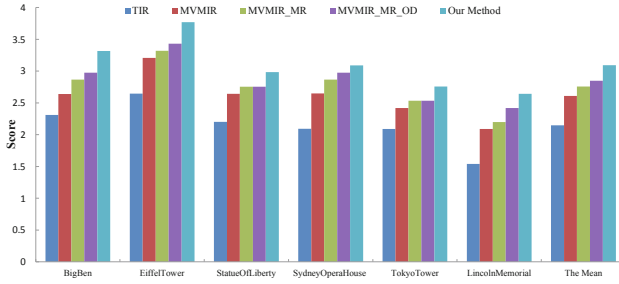


Fig. 4. Evaluation on LHV

Table 3. Illustration of visualized results by LHV

Events	Images
<p>#1 19 October 1901 Alberto Santos-Dumont in his Dirigible No.6 won a 10,000-franc prize offered by Henri Deutsch de la Meurthe for the first person to make a flight from St Cloud to the Eiffel tower and back in less than half an hour.</p>	<p>On October 19, 1901 he... awarded the Deutsch de la Meurthe prize by flying his... around the Eiffel Tower... July 13, 1901: Santos-Dumont Flies Around Eiffel Tower Brazilian aviation pioneer Alberto Santos-Dumont... No. 6 around the Eiffel tower on his way to winning a Deutsch prize. October 19, 1901</p>
<p>#2 Upon the German occupation of Paris in 1940, the lift cables were cut by the French so that Adolf Hitler would have to climb the steps to the summit.</p>	<p>Adolf Hitler posed for the... with... Adolf Hitler... June 23, 1940 1940 Adolf Hitler...</p>
<p>#3 1987 A.J. Hackett made one of his first bungee jumps from the top of the Eiffel Tower, using a special cord he had helped develop. Hackett was arrested by the Paris police upon reaching the ground</p>	<p>1987-06 This is the photo that launched a... LEGENDARY LEAP: AJ Hackett bungee jumping off the Eiffel Tower in France in 1987 Bungee jumping... New Zealand and 1987 when its... A. J. Hackett... made this... jump... a modern variant of a sport... who also ran high wooden platforms with vines tied to their sides.</p>
<p>#4 New Year's Eve 1999 The Eiffel Tower played host to Paris's Millennium Celebration. On this occasion, flashing lights and four high-power searchlights were installed on the tower, and fireworks were set off all over it</p>	<p>Tags: New Years 2000 millennium Paris celebration Tags: New Years 2000 millennium Paris celebration Tags: New Years 2000 millennium Paris celebration Title: Looking the other way Title: Fireworks everywhere Title: We had a good view Upload Time: Dec.31,1999 Upload Time: Dec.31,1999 Upload Time: Dec.31,1999</p>
<p>#5 2004 The Eiffel Tower began hosting an ice skating rink on the first floor each winter</p>	<p>Tags: eiffel tower ice rink skating Paris winter Tags: eiffel tower ice rink skating Paris winter Tags: eiffel tower ice rink skating Paris winter Title: Eiffel Tower Ice Rink Title: Eiffel Tower Ice Rink Title: Ice skating rink ON the Eiffel Tower (Dorelville) Upload Time: Dec.10,2004 Upload Time: Dec.10,2004 Upload Time: Dec.29,2004</p>
<p>#6 In January 2007, the multi-Michelin star chef Alain Ducasse was Brought in to run Jules Verne</p>	<p>Star chef Ducasse opens new Eiffel Tower restaurant Tags: eiffel tower Paris architecture Tags: eiffel tower Paris architecture Tags: eiffel tower Paris architecture Title: In front of the Jules Verne Restaurant Title: In front of the Jules Verne Restaurant Title: In front of the Jules Verne Restaurant Upload Time: Sep.26,2007 Upload Time: Sep.26,2007 Upload Time: Sep.26,2007</p>

users participating in the study. They are asked to score each obtained image with: 1) 0: if it is not relevant to the landmark; 2) 0.33: if it is not relevant to a landmark-based event; 3) 0.66: if it is not diverse with the previous ones from the same event; 4) 1: if participant is satisfied with it. The average scores of all the events for each landmark in baselines and our scheme are shown

Fig. 4. y-axes represents the average score of all events for each landmark. Since we score the top five images for each event, hence the average score of the event for each landmark is in the range $[0, 5]$. We can see that compared to the baselines, LHV achieves the best performance. This indicates that users prefer relevant and diverse images for an event. We show visualized results of some events for Eiffel Tower in Fig.3. It is worthy of noting that there are still a very few extracted events have not relevant images. We consider all sentences with temporal information as events. Although some sentences have the temporal information, they do not represent a specific event according to the definition of the event. Take one event as an example, “In 2008, a survey of 2,000 people found that the tower was the most popular landmark in the United Kingdom”. This sentence does not have corresponding images in dataset. In addition, some noisy and inaccurate tags from Flickr can also lead to the mismatch between events and relevant images.

5 Conclusions

In this paper, we have presented a new framework LHV for exploring relevant and diverse images for landmark historical events from three main knowledge sources. In our framework, we firstly extract the event list from Wikipedia and then retrieve and rerank images from image sources using these events to provide users with various results. The experiments have demonstrated the effectiveness of our approach.

Our future work is along this research direction as follows: (1) We plan to conduct experiments on more landmarks to enhance the robustness of LHV. (2) Images for some events from Wikipedia are provided, hence we plan to use these images and corresponding events to retrieve more related images from other image sources. (3) Currently, we consider landmark history visualization from three sources: Wikipedia, Flickr and Google images. We may extend our framework by incorporating more heterogeneous sources to give more satisfactory results in landmark history visualization.

Acknowledgement. This work was supported by National Program on Key Basic Research Project (973 Program, Project No. 2012CB316304), the National Natural Science Foundation of China (Grant No. 61201374, 90920303, 61003161), China Postdoctoral Science Foundation (Grant No. 2011M500430).

References

1. Allan, J., Carbonell, J., Doddington, G., Yamron, J., Yang, Y.: Topic detection and tracking pilot study final report. In: Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop (1998)
2. Bhole, A., Fortuna, B., Grobelnik, M., Mladenic, D.: Extracting named entities and relating them over time based on wikipedia. *Informatica* 31(4), 463 (2007)
3. Chasin, R.: Event and temporal information extraction towards timelines of wikipedia articles. *Simile*, 1–9 (2010)

4. He, J., Li, M., Zhang, H., Tong, H., Zhang, C.: Manifold-ranking based image retrieval. In: Proceedings of the 12th Annual ACM International Conference on Multimedia, pp. 9–16 (2004)
5. Hong, Y., Zhang, J., Ma, B., Yao, J., Zhou, G., Zhu, Q.: Using cross-entity inference to improve event extraction. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, vol. 1, pp. 1127–1136 (2011)
6. Morioka, N., Wang, J.: Robust visual reranking via sparsity and ranking constraints. In: Proceedings of the 19th ACM International Conference on Multimedia, pp. 533–542 (2011)
7. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition* 29(1), 51–59 (1996)
8. openNLP, <http://opennlp.apache.org/>
9. Strötgen, J., Gertz, M.: Heideitime: High quality rule-based extraction and normalization of temporal expressions. In: Proceedings of the 5th International Workshop on Semantic Evaluation, pp. 321–324 (2010)
10. Strötgen, J., Gertz, M., Junghans, C.: An event-centric model for multilingual document similarity. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information, pp. 953–962 (2011)
11. TimeML, <http://www.timeml.org/>
12. Torralba, A., Murphy, K., Freeman, W., Rubin, M.: Context-based vision system for place and object recognition. In: Proceedings of IEEE International Conference on Computer Vision, pp. 273–280 (2003)
13. van Leuken, R., Garcia, L., Olivares, X., van Zwol, R.: Visual diversification of image search results. In: Proceedings of the 18th International Conference on World Wide Web, pp. 341–350 (2009)
14. Wang, M., Yang, K., Hua, X., Zhang, H.: Towards a relevant and diverse search of social images. *IEEE Transactions on Multimedia* 12(8), 829–842 (2010)
15. Zhou, D., Weston, J., Gretton, A., Bousquet, O., Schölkopf, B.: Ranking on data manifolds. In: Advances in Neural Information Processing Systems, vol. 16, pp. 169–176 (2004)
16. Zhu, J., Hoi, S., Lyu, M., Yan, S.: Near-duplicate keyframe retrieval by nonrigid image matching. In: Proceedings of the 16th ACM International Conference on Multimedia, pp. 41–50 (2008)

Discovering Latent Clusters from Geotagged Beach Images

Yang Wang¹ and Liangliang Cao²

¹ Department of Computer Science, University of Manitoba, Canada

ywang@cs.umanitoba.ca

² IBM T.J.Watson Research Center, USA

liangliang.cao@us.ibm.com

Abstract. This paper studies the problem of estimating geographical locations of images. To build reliable geographical estimators, an important question is to find distinguishable geographical clusters in the world. Those clusters cover general geographical regions and are not limited to landmarks. The geographical clusters provide more training samples and hence lead to better recognition accuracy. Previous approaches build geographical clusters using heuristics or arbitrary map grids, and cannot guarantee the effectiveness of the geographical clusters. This paper develops a new framework for geographical cluster estimation, and employs latent variables to estimate the geographical clusters. To solve this problem, this paper employs the recent progress in object detection, and builds an efficient solver to find the latent clusters. The results on beach datasets validate the success of our method.

1 Introduction

Geotagged images are receiving more and more research attentions in recent years. A geotagged image is associated with a two dimensional vector, latitude and longitude, representing a unique location on the Earth. The goal of this paper is to use the visual information to estimate the geographical locations even when they are not provided. As evidenced by the success of Google Earth, there is great need for such geographic information among the mass. Many web users have high interests on not only the places they live but also other interesting places around the world. Geographic annotation is also desirable when reviewing the travel and vacation images. For example, when a user becomes interested in a nice photo, he or she may want to know where exactly it is. Moreover, if a user plans to visit a place, he or she may want to find out the points of interest nearby. Recent studies suggest that geo-tags expand the context that can be employed for image content analysis by adding extra information about the subject or environment of the image.

Estimating the geolocation of images is not an easy task. As the earlier work shown in [10] [6], only a quarter of the test images can be located subject to a rough region (approximately 750 km) near their true location. At the metropolitan scale, visual feature based annotations perform no better than chance.

As argued by [4], it is difficult to estimate the exact location at which a photo was taken. Instead, the work in [4] proposes to estimate only the coarse location in terms of geographical clusters. The goal of this paper is to find meaningful geographical clusters corresponding to different geographical regions. The use of geographical clusters can provide group wisdom for trip planning and photo organization applications. It also gathers more training samples to build more reliable classifiers.

In this paper, we focus on estimating rough geo-locations of images in terms of their geographical clusters. In particular, we use beach images in our experiments. Note that our problem is different from that of landmark recognition [23]. A landmark usually corresponds one view or one subject with a unique appearance, while a beach scene may contain a lot of clues including water, boats, people dresses, buildings and plants. Moreover, a landmark is usually limited to a point on the earth, while a beach usually covers a region. It is often inaccurate and also unnecessary to estimate the exact GPS coordinate and we only need to estimate a coarse location for a beach image.

Finding geographical clusters can lead to many applications. If we can correctly assign geolocations to image, we will be able to produce tourist maps using geographical annotation techniques [5]. We can also compare the distribution of different topics, such as cars, food, or landscapes in the world [20]. However, in practice, it is not easy to find meaningful geographical clusters. Country borders that separate the geographical regions are too coarse for large countries but too fine for small ones. [3] proposed to initialize meaningful geographical clusters by spatial clustering refine the cluster by post processing. In this paper, we will discuss a new method to find the geographical clusters using an efficient latent SVM learning.

2 Previous Work

Geographical annotation provides a rich source of information which can link millions of images based on the similarity of their geographical locations. There have been a growing body of work in visual research community investigating geographical information for image understanding [15] [1] [4] [21] [11] [22] [12] [14] [16] [13] [17] [20] [2]. Many applications are motivated by Jim Gray's idea to build a personal Memex which can record everything a person sees and hears, and quickly retrieve any item on request. Moreover, It is more interesting to aggregate information from a large number of users, so that group wisdom can be mined from these media. As suggested by [2], if we know a number of user favored images, we can provide effective tourism recommendation under the premise "If you like this picture, you will also like these places". However, such a personal Memex requires a huge amount of geo-tagged information, which is still not practical given the fact that 99% of Flickr photos do not have related geographical information associated with them.

To address the challenges, one group of research work is devoted to estimating the geographical information from general images. Hays and Efros [10] are

among the first to consider the problem of estimating the location of a single image using only its visual content. They collect millions of geo-tagged Flickr images. Using a comprehensive set of visual features, they employ nearest neighbor search in the reference set to locate the image. Motivated by [10], Gallagher et al. [9] incorporate textual tags to estimate the geographical locations of images. Their results show that textual tags perform better than visual content and the combination of textual and visual information performs better than either alone. Cao et al. [4] also recognize the effectiveness of tags in estimating the geolocations. They propose a novel model named logistic canonical correlation regression which explores the canonical correlations between geographical locations, visual content and community tags. Unlike [10], they argue that it is difficult to estimate the exact location at which a photo was taken and propose to estimate only the coarse location. Similarly, Crandall et al. [6] only estimate the approximate location of a novel photo. Using SVM classifiers, a novel image is geolocated by assigning it to the best cluster based on its visual content and annotations. In a recent research work [23] supported by Google, Zhen et al. built a web-scale landmark recognition engine named “Tour the world” using 20 million GPS-tagged photos of landmarks together with online tour guide web pages. The experiments demonstrate that the engine can deliver satisfactory recognition performance with high efficiency.

Despite of these research efforts, recognizing the location of a non-landmark image reliably is still an open question. For those non-landmark locations, visual information based classifiers only perform comparable to chance. A recent study [3] propose to discover “geographical clusters” to build classifiers. The use of geographical clusters benefits the problem of localization in two aspects: On the training stage, geographical clusters provide more training samples and hence lead to better recognition accuracy; on the testing stage, estimation of the most possible region for each query photo will be relatively easier than the estimation of exact GPS coordinates, while the information of geographical cluster will be good enough for trip planing and photo organization applications. However, the geographical clusters in [3] are discovered by refined mean-shift clusters, which are not representative enough for visual recognition. In this paper, we aim to develop a more principled approach to find geographical clusters.

This paper is motivated by the some recent progress in object detection [8] and max-margin clustering [18]. In the object detection method of [8], the locations of object parts are unknown, and are treated as latent (hidden) variables in a learning framework called the *latent SVM*. A latent SVM is an extension of regular SVMs to handle latent variables. A latent SVM is semi-convex and the training problem becomes convex once latent information is specified for the positive examples. This leads to an iterative training algorithm that alternates between fixing latent values for positive examples and optimizing the latent SVM objective function. Similar ideas can also be found in [18] which finds maximum margin hyperplanes through data. In this paper, we treat the geographical clusters of training images as hidden labels, and develop a principled learning method that recognizes the geographical clusters of images.

3 Our Approach

The work in [3] finds geographical clusters by clustering the GPS coordinate vectors of training images. Then a SVM classifier based on image features is learned for each cluster. For a new test image, the SVM classifier can be used to assign this image to a corresponding cluster based on its image feature. A limitation of this approach is that clustering and SVM learning are treated as two independent tasks. However, we believe these two tasks should be coupled together. In this paper, we propose a new approach that considers image clustering and model learning in a single unified framework.

3.1 Geo-location Regularized Clustering

Our method is based on the max-margin clustering (MMC) [18]. Naively applying MMC to our dataset is troublesome, since MMC is a generic clustering algorithm and does not take into account of the geo-location information of the data. We propose an extension of MMC that clusters training images so that images in the same cluster are both visually similar and have close GPS locations.

We assume that we are given a training dataset with N instances. Each instance is in the form of (x_i, y_i) , where x_i is the i -th image, and y_i is its corresponding geo-location. Our goal is to cluster the training images into C groups in some sensible manner. We would also like to have a discriminative model that can assign an unseen image to one of the clusters. If we ignore the geo-location information y_i in the training data and only consider the image feature x_i , we can use standard clustering algorithms to partition the training images into C clusters. But now the challenge is how to incorporate the GPS location information into the clustering process.

Let us assume that the number of clusters is known to be C . Clustering the training data is equivalent to assigning a binary vector z_i to each image x_i . Here z_i is a vector of length C , where its c -th component z_{ic} is defined as:

$$z_{ic} = \begin{cases} 1 & \text{if } x_i \text{ belongs to cluster } c \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Note that if z_i is observed on training data, we can use this information to learn a multi-class SVM classifier to assign the cluster membership of an unseen image by solving the following optimization problem:

$$\mathcal{P}(w^*) = \min_{w, \xi} \frac{1}{2} \|w\|^2 + C_1 \sum_i \xi_i \quad (2a)$$

$$\text{s.t. } w^\top \phi(x_i, z_i) - w^\top \phi(x_i, z) \geq \Delta(z_i, z) - \xi_i, \quad \forall i, \forall z \quad (2b)$$

where w and $\phi(x_i, z_i)$ is a feature vector, ξ_i is the slack variable for handling soft margins in SVM classifiers.

Now since z_i is not observed, we need to simultaneously partition the training data into C groups and learn the multi-class SVM. Using the same reasoning of unsupervised SVM [18,19], we can try to solve the following optimization problem:

$$\mathcal{P}(w^*, \{z_i\} : \forall i) = \min_{w, \xi} \min_{\{z_i\}} \frac{1}{2} \|w\|^2 + C_1 \sum_i \xi_i \quad (3a)$$

$$\text{s.t. } w^\top \phi(x_i, z_i) - w^\top \phi(x_i, z) \geq \Delta(z_i, z) - \xi_i, \quad \forall i, \forall z \quad (3b)$$

Note that in Eq. 3, we need to optimize over the variables $\{z_i\}$, since they are unknown on the training data. The optimization problem in Eq. 3 tries to find $\{z_i\}$ so that the resultant SVM has the maximum margin (please refer to [18,19] for details).

Unfortunately, without additional constraints or regularization, Eq. 3 has a degenerate solution. Basically we can assign all training data to the same cluster and learn w to achieve arbitrarily large margin. In [18,19], this problem is addressed by adding a constraint that tries to make sure that the clusters are balanced.

For our application, we have the additional information (i.e. GPS locations) in addition to images. In the following, we will use this additional information to regularize Eq. 3. Intuitively, we would like the clusters to have the following property. If two images are close in terms of their geo-locations, they are more likely to be in the same cluster. One natural way to formalize this intuition is to solve the following optimization problem:

$$\mathcal{P}(w^*, \{z_i\} : \forall i) = \min_{w, \xi} \min_{\{z_i\}} \frac{1}{2} \|w\|^2 + C_1 \sum_i \xi_i \quad (4a)$$

$$+ C_2 \sum_i \sum_j (-|z_i - z_j| d_{ij}) \quad (4b)$$

$$\text{s.t. } w^\top \phi(x_i, z_i) - w^\top \phi(x_i, z) \geq \Delta(z_i, z) - \xi_i, \quad \forall i, \forall z \quad (4c)$$

where d_{ij} is the distance of two images x_i and x_j in terms of their geo-locations (which can be obtained from y_i and y_j).

Note that $|z_i - z_j| = 0$ if i and j are in the same cluster. So Eq. 4b will try to make the distance (in terms of GPS locations) between images in different clusters to be large.

The optimization problem in Eq. 4 can be solved using an iterative approach:

- Fix $\{z_i\}_{i=1}^N$, optimize over w and ξ .
- Fix w and ξ , optimize over $\{z_i\}_{i=1}^N$.

The first step of this iterative approach is straightforward since it is equivalent to solving a standard multi-class SVM problem. The second step is more

challenging. It involves solving a combinatorial problem which can be shown to be NP-hard.

One possible solution is to use linear program relaxation to get an approximate solution. But the resultant linear program is still too large to be practical. In the following section, we introduce a new formulation that is more amenable to efficient algorithms.

3.2 More Efficient Formulation

The main observation that enables our new formulation is the following. Suppose we know the cluster centers $\{g_c\}_{c=1}^C$ (in term of geo-locations), a natural way to solve our problem is to use the following optimization:

$$\mathcal{P}(w^*, \{z_i\} : \forall i) = \min_w \min_{\xi} \frac{1}{2} \|w\|^2 + C_1 \sum_i \xi_i \quad (5a)$$

$$+ C_2 \sum_i \sum_c (z_{ic} \|y_i - g_c\|^2) \quad (5b)$$

$$\text{s.t. } w^\top \phi(x_i, z_i) - w^\top \phi(x_i, z) \geq \Delta(z_i, z) - \xi_i, \quad \forall i, \forall z \quad (5c)$$

Note that Eq. 5b computes the distance (in term of geo-locations) between images and their corresponding cluster centers. When those cluster centers are known, the optimal clustering is obtained by choosing cluster membership that minimizes this distance (i.e. minimizing over $\{z_i\}$).

Now the challenge is that the cluster centers $\{g_c\}$ are also unknown. Using the same reasoning in Sec. 3.1, we propose to treat the cluster centers as yet another set of latent variables in the formulation and use the following iterative method to solve it:

- Fix $\{z_i\}_{i=1}^N$ and $\{g_c\}_{c=1}^C$, optimize over w and ξ : this step is equivalent to solving a regular multi-class SVM. We use liblinear [7] for it.
- Fix w , ξ and $\{z_i\}_{i=1}^N$, optimize over $\{g_c\}_{c=1}^C$: it is easy to show that if we use the l_2 distance, the optimal value of the c -th cluster center g_c is the average of the geo-locations of images assigned (based on $\{z_i\}$) to this cluster.
- Fix w , ξ and $\{g_c\}_{c=1}^C$, optimize over $\{z_i\}_{i=1}^N$: it is easy to show this step is a linear assignment problem.

4 Experiments

We test our approach on a dataset containing images downloaded from Flickr with the tags of “beach” or “coast”. Each photo is associated with a two-dimensional GPS coordinate vector. Similar to [3], we use 34558 images for training and 1185 images for testing. We use GIST features to represent images.

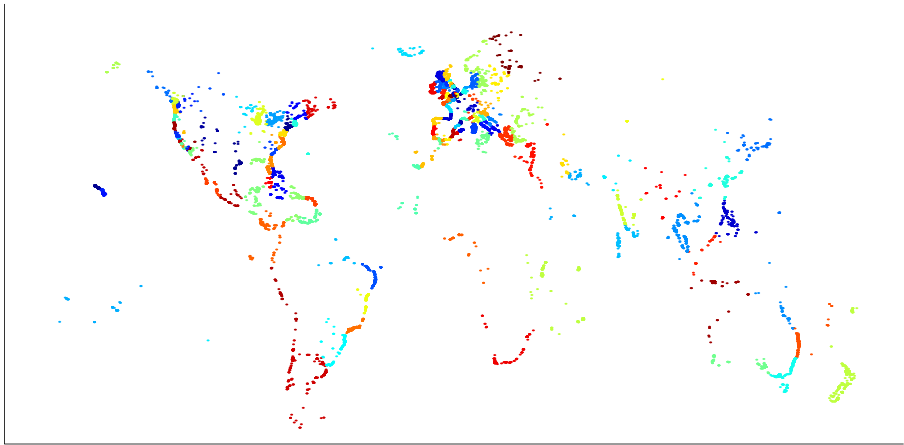


Fig. 1. Visualization of clustering training images using our method. Each color represents a different cluster.

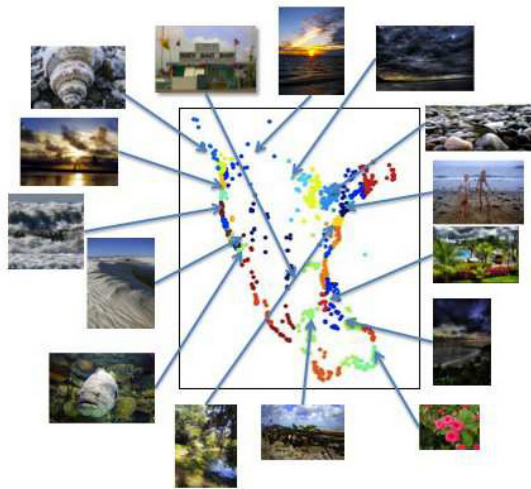


Fig. 2. Visualization of representative images for North America

In Fig. 1, we plot the distribution of training images and their clusters in roughly different colors. In Figs. 2 3 4, we visualize some representative images in some clusters.

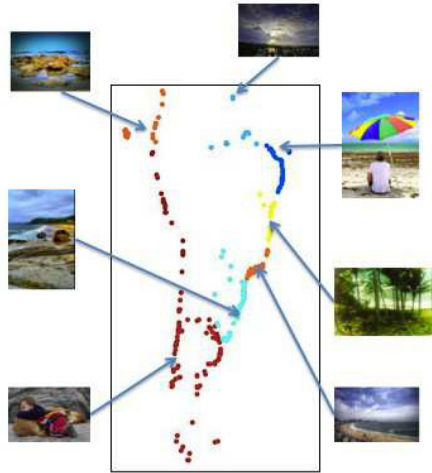


Fig. 3. Visualization of representative images for South America

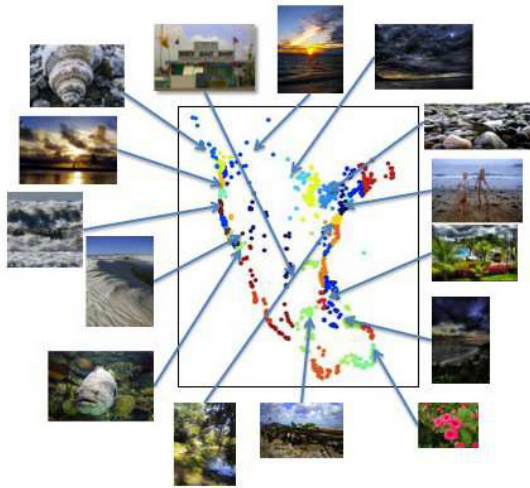


Fig. 4. Visualization of representative images for Asia

5 Conclusion

We have introduced a new framework for geographical cluster estimation. Our approach treats the geographical cluster of an image as a latent variable. Our method jointly clusters training images and learns discriminative classifiers for each cluster in a single framework.

Acknowledgement. Yang Wang is supported by a start-up grant from the University of Manitoba.

References

1. Agarwal, M., Konolige, K.: Real-time localization in outdoor environments using stereo vision and inexpensive GPS. In: International Conference on Pattern Recognition (2006)
2. Cao, L., Luo, J., Gallagher, A., Jin, X., Han, J., Huang, T.: A worldwide tourism recommendation system based on geotagged web photos. In: International Conference on Acoustics, Speech, and Signal Processing, ICASSP (2010)
3. Cao, L., Smith, J., Wen, Z., Yin, Z., Jin, X., Han, J.: BlueFinder: Estimate where a beach photo was taken. In: WWW (2012)
4. Cao, L., Yu, J., Luo, J., Huang, T.: Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In: Proceedings of the Seventeenth ACM International Conference on Multimedia, pp. 125–134 (2009)
5. Chen, W., Battestini, A., Gelfand, N., Setlur, V.: Visual summaries of popular landmarks from community photo collections. In: ACM International Conference on Multimedia, pp. 789–792 (2009)
6. Crandall, D., Backstrom, L., Huttenlocher, D., Kleinberg, J.: Mapping the world’s photos. In: International Conference on World Wide Web, pp. 761–770 (2009)
7. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A Library for Large Linear Classification. JMLR (2008)
8. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32(9), 1627–1645 (2010)
9. Gallagher, A., Joshi, D., Yu, J., Luo, J.: Geo-location Inference from Image Content and User Tags
10. Hays, J., Efros, A.A.: Im2gps: estimating geographic information from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)
11. Joshi, D., Luo, J.: Inferring generic places based on visual content and bag of geotags. In: ACM Conference on Content-based Image and Video Retrieval (2008)
12. 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: ACM Conference on Multimedia (2007)
13. Luo, J., Yu, J., Joshi, D., Hao, W.: Event recognition: viewing the world with a third eye. In: ACM International Conference on Multimedia, pp. 1071–1080 (2008)
14. Naaman, M.: Leveraging geo-referenced digital photographs. PhD thesis, Stanford University (2005)
15. Naaman, M., Song, Y., Paepcke, A., Garcia-Molina, H.: Automatic organization for digital photographs with geographic coordinates. In: International Conference on Digital Libraries, vol. 7, pp. 53–62 (2004)
16. Quack, T., Leibe, B., Van Gool, L.: World-scale mining of objects and events from community photo collections. In: ACM Conference on Image and Video Retrieval, pp. 47–56 (2008)
17. Schindler, G., Krishnamurthy, P., Lubliner, R., Liu, Y., Dellaert, F.: Detecting and matching repeated patterns for automatic geo-tagging in urban environments. In: IEEE Conference on Computer Vision and Pattern Recognition (2008)

18. Xu, L., Neufeldand, J., Larson, B., Schuurmans, D.: Maximum margin clustering. In: Saul, L.K., Weiss, Y., Bottou, L. (eds.) *Advances in Neural Information Processing Systems*, vol. 17, pp. 1537–1544. MIT Press, Cambridge (2005)
19. Xu, L., Wilkinson, D., Southey, F., Schuurmans, D.: Discriminative unsupervised learning of structured predictors. In: *Proceedings of the 23th International Conference on Machine Learning* (2006)
20. Yin, Z., Cao, L., Han, J., Zhai, C., Huang, T.: Geographical topic discovery and comparison. In: *Proceedings of the 20th International Conference on World Wide Web*, pp. 247–256. ACM (2011)
21. Yu, J., Luo, J.: Leveraging probabilistic season and location context models for scene understanding. In: *International Conference on Content-based Image and Video Retrieval*, pp. 169–178 (2008)
22. Yuan, J., Luo, J., Wu, Y.: Mining compositional features for boosting. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2008)
23. Zheng, Y., Zhao, M., Song, Y., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T., Neven, H.: Tour the World: building a web-scale landmark recognition engine. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2009)

An Error Resilient Depth Map Coding Scheme Using Adaptive Wyner-Ziv Frame

Xiangkai Liu¹, Qiang Peng¹, Xiao Wu¹, Lei Zhang¹,
Xu Xia¹, and Lingyu Duan²

¹ School of Information Science & Technology,
Southwest Jiaotong University, Chengdu, 610031, China

² The Institute of Digital Media, School of EE&CS,
Peking University, Beijing, 100871, China
{liuxiangkai,xiayu}@gmail.com,
{qpeng,wuxiaohk,zhanglei}@home.swjtu.edu.cn,
lingyu@pku.edu.cn

Abstract. Depth information is one of the most important parameters in three dimensional videos (3DV). While transmitted through error-prone networks, the distortion in depth map due to packet loss will lead to a geometric error during the process of Depth Image Based Rendering (DIBR) and affect the rendered video quality. In this paper, we propose an error resilient depth map coding scheme using adaptive Wyner-Ziv (WZ) frames. For each depth frame, whether to be encoded into a Wyner-Ziv frame is decided by a joint source-channel R-D optimization (JSC-RDO) algorithm. JSC-RDO involves in the end-to-end distortion model of depth coding and the estimation of expected rate and distortion of WZ coding. Motion information of the corresponding color video are used to correct the depth error and generate the side information for WZ coding. The Lagrange multiplier used in JSC-RDO was derived taking into account both the packet-loss environment and the rendered view distortion. Experimental results show that the proposed error resilient scheme achieves a better overall R-D performance than existing schemes.

Keywords: 3DV, depth map, error resilient, Wyner-Ziv.

1 Introduction

Multi-view video plus depth (MVD) is a new video format for 3D video discussed in MPEG-3DV [1], it contains multi-view color video and depth map. By using depth map, it is possible to efficiently render virtual views in decoder side based on Depth Image Based Rendering (DIBR) technique [2]. However in transmission, if some of these depth information are lost due to network errors, it will lead to degraded rendered video quality at the receiver, as the distortion in depth map will induce a geometric error in the process of DIBR.

Robust video coding over the existing packet-switched networks is an important topic of research due to the problems caused by packet loss. These transmission problems have inspired several feasible solutions. One category of solutions

focuses on the error concealment tools devised in the decoder, another category is based on the error resilient coding tools devised in the encoder. While error concealment approaches try to estimate the lost blocks in a video frame, error resilient approaches try to generate more robust encoded bit-stream. There are several techniques for error concealment of depth coding that have been developed in the past [3][4][5]. In [3], the motion correlation between the color and depth video streams are exploited for error concealment. In [4][5], the internal characteristics of the macroblock in the depth map was investigated and used to recover accurately the lost motion vector for the corrupted blocks. However, all these schemes only provide an error correction algorithm at the decoder side, none of them were aimed at error resilient coding at the encoder side.

Recently, distributed source coding, more specifically, Wyner-Ziv (WZ) coding [6][7], emerges as a promising scheme for error-resilient video coding, which integrates the encoder-driven error resilience and the decoder-driven error concealment. These schemes attempted to terminate temporal error propagation using WZ protected frames. However, these WZ frames are inserted periodically, not adaptively, so the coding efficiency cannot be satisfactory. Also, none of these techniques were aimed at depth video coding.

In this paper, we propose an error resilient depth map coding scheme where an adaptively chosen subset of these depth frames are coded as WZ frame to prevent temporal error propagation caused by packet-loss. To decide which depth frame should be coded as WZ frame, a joint source-channel R-D optimization (JSC-RDO) algorithm is developed. It refers to three major contributions. First, the end-to-end distortion model combining an error concealment method for depth coding is proposed. Second, the Lagrange multiplier is derived taking into account both the packet-loss environment and the rendered view distortion. Finally, the expected rate and distortion of WZ coding is estimated depend on the side information generated by our depth error concealment method.

The rest of the paper is organized as follows. Section 2 formulates the generic JSC-RDO algorithm, the end-to-end distortion model with error concealment method and the optimized Lagrange multiplier is derived. In Section 3, the estimation of the rate and distortion of WZ coding is presented. Section 4 shows the simulation results, and Section 5 conclude the paper.

2 Joint Source Channel R-D Optimization for Depth Map Coding

2.1 Problem Formulation

In our proposed depth coding scheme, two candidate frame type are available for each frame: conventional I/P frame and WZ frame. The best coding type t of frame n can be selected as the one having the minimum coding cost $J(n, t)$:

$$\begin{aligned} J(n, t) &= D(n, t) + \lambda R(n, t) \\ &= D_{render}(D_{depth}(n, t)) + \lambda R(n, t) \end{aligned} \quad (1)$$

where n is the frame number and t indicates the frame type. The Lagrange multiplier λ finds the tradeoff between the distortion $D(n, t)$ and the encoding rate $R(n, t)$. $D_{depth}(n, t)$ denotes the depth map distortion, D_{render} is the distortion in rendered view expressed as a function of the depth distortion. Because depth maps are used to help view rendering process but will not be directly displayed, we use D_{render} as the distortion metric in our R-D optimization process. For I/P frame, the Lagrangian cost is

$$\begin{aligned} J(n, I/P) &= D_{render}(D_{depth_sc}(n, I/P)) + \lambda R_s(n, I/P) \\ &= D_{render}(D_{depth_s}(n, I/P) + D_{depth_c}(n, I/P)) + \lambda R_s(n, I/P) \end{aligned} \quad (2)$$

where $D_{depth_sc}(n, I/P)$ denotes the end-to-end distortion of depth coding, say joint source-channel distortion. $D_{depth_s}(n, I/P)$ and $R_s(n, I/P)$ is the source coding distortion and bit-rate, respectively. $D_{depth_c}(n, I/P)$ is the channel transmission distortion. For WZ frame, the Lagrangian cost is

$$J(n, WZ) = D_{render}(D_{depth_wz}(n)) + \lambda R_{wz}(n) \quad (3)$$

where $D_{depth_wz}(n)$ denotes the reconstruction distortion of the depth map after WZ decoding, and $R_{wz}(n)$ is the WZ coding bit-rate. For a depth frame n , after coded with a standardized H.264/AVC engine, we compute the cost $J(n, I/P)$ and $J(n, WZ)$, if $J(n, WZ) < J(n, I/P)$, frame n will be coded as WZ frame, otherwise, frame n will be coded as I/P frame. Considering the computation complexity, we only operate our JSC-RDO algorithm at frame level and the MB mode selection process in I/P frame coding is the same as the H.264/AVC standard.

Since the source coding bit-rate $R_s(n, I/P)$ can be found during the coding process, we will focus on the problem about how to estimate $D_{depth_sc}(n, I/P)$, $D_{depth_wz}(n)$ and $R_{wz}(n)$, the calculation of D_{render} will also be introduced.

2.2 End-To-End Distortion Model

Firstly, we define some notations used in the derivation of the proposed end-to-end distortion model. For pixel i in frame n that references pixel j in frame ref , let f_n^i be the original value, and let \hat{f}_n^i and \tilde{f}_n^i be the reconstructed values in the encoder and decoder, respectively. Let \hat{r}_n^i be the reconstructed residue in the encoder, i.e., $\hat{f}_n^i = \hat{f}_{ref}^j + \hat{r}_n^i$. When the current pixel is lost in the decoder, it will be concealed by pixel k in frame n' . Suppose the transmission error rate is p . Then, we can represent \tilde{f}_n^i as

$$\tilde{f}_n^i = \begin{cases} \hat{f}_{ref}^j + \hat{r}_n^i & w.p. \quad 1 - p \\ \hat{f}_{n'}^k & w.p. \quad p \end{cases} \quad (4)$$

Let $d_{depth_sc}(n, i)$ be the expected end-to-end distortion of pixel i in depth frame n , and the $D_{depth_sc}(n, I/P)$ in (3) can be obtained as

$$D_{depth_sc}(n, I/P) = \sum_{i=0}^N d_{depth_sc}(n, i) \quad (5)$$

where N denotes the frame size, and we can derive $d_{depth_sc}(n, i)$ as

$$\begin{aligned}
d_{depth_sc}(n, i) &= E\{(f_n^i - \tilde{f}_n^i)^2\} \\
&= (1-p)E\{(f_n^i - (\tilde{f}_{ref}^j + \tilde{r}_n^i))^2\} + pE\{(f_n^i - \tilde{f}_{n'}^k)^2\} \\
&= (1-p)E\{(f_n^i - \tilde{f}_n^i)^2\} + (1-p)E\{(\tilde{f}_{ref}^j - \tilde{f}_{ref}^j)^2\} \\
&\quad + pE\{(f_n^i - \tilde{f}_{n'}^k)^2\} \\
&= (1-p)d_{depth_s}(n, i) + (1-p)d_{depth_ep}(ref, j) \\
&\quad + pd_{depth_ec}(n, i)
\end{aligned} \tag{6}$$

where $d_{depth_s}(n, i)$ denotes the pixel source distortion, $d_{depth_ep}(ref, j)$ denotes the error-propagated distortion from the reference frame, and $d_{depth_ec}(n, i)$ denotes the error-concealment distortion. Since $d_{depth_s}(n, i)$ can be found in the encoding process, we will focus on the problem about how to obtain $d_{depth_ec}(n, i)$ and $d_{depth_ep}(ref, j)$.

Before calculate $d_{depth_ec}(n, i)$, it is important to introduce our error concealment method, as we know the depth map is a gray scale version of the corresponding color video and there are strong correlation between their motion information, so we can use the motion vector of the corresponding color pixel to find $\tilde{f}_{n'}^k$, if the color pixel is intra coded without motion information, the depth pixel at the same place in frame $n-1$ will be copied. The error concealment method can be written as

$$\tilde{f}_{n'}^k = \begin{cases} \tilde{f}_{ref'}^{j'} & f_{color_n}^i \text{ is inter} \\ \tilde{f}_{n-1}^i & f_{color_n}^i \text{ is intra} \end{cases} \tag{7}$$

where $f_{color_n}^i$ is the corresponding color pixel of the lost depth one, $\tilde{f}_{ref'}^{j'}$ is the depth pixel whose frame number ref' and pixel number j' is copied from the reference pixel of $f_{color_n}^i$. Now we can derive $d_{depth_ec}(n, i)$ as

$$\begin{aligned}
d_{depth_ec}(n, i) &= E\{(f_n^i - \tilde{f}_{n'}^k)^2\} \\
&= E\{(f_n^i - \tilde{f}_{n'}^k)^2\} + E\{(\tilde{f}_{n'}^k - \tilde{f}_{n'}^k)^2\} \\
&= E\{(f_n^i - \tilde{f}_{n'}^k)^2\} + d_{depth_ep}(n', k)
\end{aligned} \tag{8}$$

the remained problem is to calculate $d_{depth_ep}(n', k)$ in (8) and $d_{depth_ep}(ref, j)$ in (6). Without losing the generality, we derive $d_{depth_ep}(n, i)$ as

$$\begin{aligned}
d_{depth_ep}(n, i) &= E\{(\hat{f}_n^i - \tilde{f}_n^i)^2\} \\
&= (1-p)E\{(\hat{f}_n^i - (\tilde{f}_{ref}^j + \tilde{r}_n^i))^2\} + pE\{(\hat{f}_n^i - \tilde{f}_{n'}^k)^2\} \\
&= (1-p)E\{(\hat{f}_{ref}^j - \tilde{f}_{ref}^j)^2\} + pE\{(\hat{f}_n^i - \tilde{f}_{n'}^k + \hat{f}_{n'}^k - \tilde{f}_{n'}^k)^2\} \\
&= (1-p)E\{(\hat{f}_{ref}^j - \tilde{f}_{ref}^j)^2\} + pE\{(\hat{f}_n^i - \tilde{f}_{n'}^k)^2\} \\
&\quad + pE\{(\hat{f}_{n'}^k - \tilde{f}_{n'}^k)^2\} \\
&= (1-p)d_{depth_ep}(ref, j) + pE\{(\hat{f}_n^i - \tilde{f}_{n'}^k)^2\} + pd_{depth_ep}(n', k)
\end{aligned} \tag{9}$$

we can find calculating d_{depth_ep} is a recursive process, and note that the d_{depth_ep} of the first frame can be directly derived without considering the error propagation because it is typically coded as an I frame, so the d_{depth_ep} of the following frames can be recursively calculated frame by frame.

2.3 Distortion Estimation of Rendered View

The D_{render} in (1) is the distortion in rendered view expressed as a function of the depth distortion. The distortion in depth map will cause geometry errors during the process of DIBR [8], the pixel position error $(\Delta x, \Delta y)$ in rendered view due to depth error can be calculated as follows

$$\begin{pmatrix} \Delta x \\ \Delta y \\ 1 \end{pmatrix} = \frac{\Delta Depth(x, y)}{255} \left(\frac{1}{Depth_{near}} - \frac{1}{Depth_{far}} \right) \mathbf{AR}\{\mathbf{T}' - \mathbf{T}\} \quad (10)$$

where $\Delta Depth(x, y)$ is the depth error in position (x, y) . $Depth_{near}$ and $Depth_{far}$ are the nearest and the farthest depth value in the scene. \mathbf{A} , \mathbf{R} and \mathbf{T} are respectively the intrinsic, rotation and translation matrix of the rendered view, and \mathbf{T}' belongs to the reference view.

In a DIBR system, a view can be rendered using color video and its corresponding depth map. The exact amount of distortion in the rendered view can be measured if we compare the rendered view with the ground truth. However, the ground truth may not be available since the rendered view can be generated for any arbitrary viewpoint. Instead, we propose to use the reference color video frame that belong to the same viewpoint as the depth map, which is always available since it is encoded along with the depth map. So we calculate the distortion in rendered view as

$$d_{render}(n, i) = (f_{color_n}^i - f_{color_n}^{i+\Delta P})^2 \quad (11)$$

where $d_{render}(n, i)$ is the distortion of pixel i in rendered frame n , and $f_{color_n}^i$ is the corresponding pixel in reference color frame, ΔP is the rendering position error, $f_{color_n}^{i+\Delta P}$ is obtained by moving from the position of $f_{color_n}^i$ with ΔP .

2.4 Derivation of Lagrange Multiplier

The optimal Lagrange multiplier can be selected by taking derivative of distortion and bit-rate. When this is applied to the error resilient depth map coding, it is necessary to consider the rendered view distortion and the packet-loss environment. If the distortion and the bit-rate are expressed as a function of quantization step size Q , λ can be calculated by taking the derivative of (1) and setting it to zero

$$\lambda = -\frac{dD_{render}(Q)/dQ}{dR(Q)/dQ} \quad (12)$$

as in (10), there is a linear relationship between the depth map distortion ΔD_{depth} , and the rendering position error ΔP in the rendered view

$$\Delta D_{depth} = \gamma \Delta P \quad (13)$$

γ is a scale factor determined by (10). Because the effect of geometry error on the quality of rendered view will depend on local characteristics of the video. For example, in areas of a video frame with complex textures and objects, the distortion caused by the geometry error will be significant, as different positions should have quite different pixel values. Therefore, it is necessary to link the geometry error to the rendered view distortion according to the local video characteristics

$$\Delta D_{render} = \eta \cdot \sigma^2 \Delta P \quad (14)$$

where σ^2 is the variance of the frame representing the local characteristics of the video such as texture complexity, and η is a scaling factor. Assuming high-resolution quantization, it is well known that source distortion D_{depth_s} conforms to

$$D_{depth_s}(R) = \beta \cdot 2^{-\alpha R} \quad (15)$$

where β is a constant depending on the variance of the source. Further assuming that the distortion-to-quantize relation is at sufficiently high rates, the source probability distribution can be approximated as uniform within each quantization interval Q

$$D_{depth_s}(Q) = \frac{Q^2}{12} \quad (16)$$

Combining (15) and (16), we obtain

$$R(Q) = \frac{1}{\alpha} \log_2 \left(\frac{\beta}{D_{depth_s}(Q)} \right) = \frac{1}{\alpha} \log_2 \left(\frac{\beta}{Q^2/12} \right) \quad (17)$$

According to (2)(6)(13)(14)and(16), we also obtain

$$D_{render}(Q) = \frac{1-p}{\delta \cdot \sigma^2} \frac{Q^2}{12} + (1-p)D_{ep} + pD_{ec} \quad (18)$$

where $\delta = \frac{\eta}{\gamma}$. Note that both D_{ep} and D_{ec} are independent of the quantization interval Q of the current frame. Further combining the derivatives for Q in (17) and (18), we can derive the new Lagrange multiplier as

$$\lambda = -\frac{dD(R)}{dR} = -\frac{dD}{dQ} \frac{dQ}{dR} = \frac{1-p}{\delta \cdot \sigma^2} \frac{\alpha \ln 2}{12} Q^2 = \frac{1-p}{\delta \cdot \sigma^2} \lambda_0 \quad (19)$$

where λ_0 indicates the Lagrange multiplier in H.264 standard

$$\lambda_0 = 0.85 \cdot 2^{\frac{Q-12}{3}} \quad (20)$$

3 Bit-Rate and Distortion Estimation for WZ Coding

3.1 Bit-Rate Estimation for WZ Coding

The WZ coding bit-rate R_{wz} is determined by the conditional entropy $H(X|Y)$ which is related with the correlation noise level between source X and its side information Y , i.e., the transmission distortion D_{depth_c} in the context of this paper. Since WZ coding operates on a bit-plane basis, the conditional entropy is estimated for each bit-plane. Supposing x is a coefficient from one coefficients-band of X , and y is the corresponding coefficient of the side information. The conditional entropy of the t th bit-plane of x can be computed as

$$H(b_t^x|\varepsilon) = -p(0|\varepsilon) \cdot \log_2(p(0|\varepsilon)) - p(1|\varepsilon) \cdot \log_2(p(1|\varepsilon)) \quad (21)$$

where ε represents available information at the decoder side, specifically, previously decoded bit-planes $\{b_{N-1}^x, \dots, b_{t+1}^x\}$ and side information y , N is the number of bit-planes and $t \in [0, N-1]$. $p(b_t^x|\varepsilon)$ is the conditional probability of b_t^x given ε . To obtain $p(b_t^x|\varepsilon)$, we adopt the *a priori* probability partition algorithm [6], where the range of conditional probability distribution $p(b_t^x|\varepsilon)$ is equally divided into two parts whose areas represent the probability that the coming bit b_t^x equals to 0 and 1, respectively. For each b_t^x , the $p(b_t^x|\varepsilon)$ can be obtained by

$$p(b_t^x = 0|\varepsilon) = \frac{(2^t - 1) + \sum_{j=t+1}^N b_j^x 2^j}{\sum_{j=t+1}^N b_j^x 2^j} p(x|y) dx \quad (22)$$

$$p(b_t^x = 1|\varepsilon) = \frac{(2^{t+1} - 1) + \sum_{j=t+1}^N b_j^x 2^j}{2^t + \sum_{j=t+1}^N b_j^x 2^j} p(x|y) dx \quad (23)$$

Note that for AC coefficients, the first bit-plane is the sign bit-plane. To calculate the conditional entropy, $p(x|y)$ is the most important parameter which can be derived from $p(e)$, i.e., $p(x|y) = p(y + e|y)$, $e = x - y$ is the transmission error. In previous work [10], the $p(e)$ was presented using a Laplacian distribution

$$p(e) = \frac{\alpha}{2} \exp(-\alpha e) \quad (24)$$

α is the Laplacian distribution parameter defined by

$$\alpha = \sqrt{\frac{2}{\sigma^2}} \quad (25)$$

where σ^2 is the variance of the residual between the source frame X and side information Y . To obtain σ^2 , we firstly compute the residual frame between X and Y , then calculate its DCT coefficients frame T

$$T = DCT(X - Y) \quad (26)$$

For each DCT-band b , there is a set of coefficients T_b and a variance σ_b , and we compute σ_b^2 as

$$\sigma_b^2 = E(T_b^2) - E(T_b)^2 \quad (27)$$

Now the remaining problem is how to generate the side information Y at the encoder side. In the packet-loss environment, Y consists of two parts

$$Y = Y_{error_free} + Y_{error}(D_{depth_c}) \quad (28)$$

where Y_{error_free} denotes the side information generated in error free environment and Y_{error} is related to the channel transmission distortion D_{depth_c} . However, as the actual transmission error is not available at encoder, the exact value of (28) can not be calculated straightforward. In this paper, the side information Y at encoder is approximated with a uniformly distributed transmission error pattern, given the specific error pattern, Y can be generated using the error concealment method in (7). Finally, the estimated bit-rate R_{wz} can be calculated as

$$R_{wz} = c \cdot \sum_{j=0}^{K-1} \sum_{i=0}^{M-1} \sum_{t=0}^{N-1} H(b_t^{x_{ij}} | \varepsilon) \quad c \geq 1 \quad (29)$$

where K is the number of coefficient-band, M is the number of coefficient in each band, and N is the bit-plane number in coefficient-band j . The calibration coefficient c depends on the efficiency of the specific WZ decoder, we set $c = 1.3$ to avoid decoding failure when no feedback channel is available.

3.2 Distortion Estimation for WZ Coding

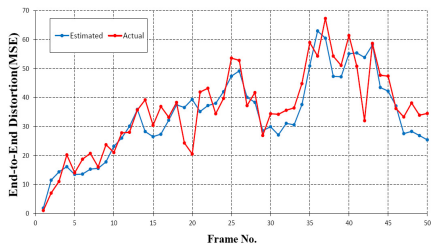
The distortion of WZ coding comes from the process of reconstruction. When we simulate the state-of-the-art MMSE inverse quantization [11] at the encoder, the distortion D_{wz} can be estimated as

$$D_{wz} = (X - \tilde{X})^2 \quad (30)$$

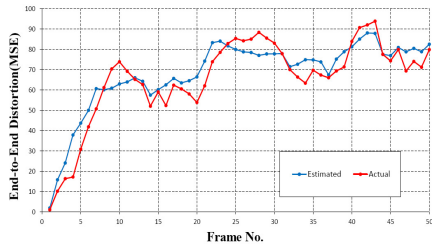
where \tilde{X} is the reconstructed frame obtained by the simulated decoding process at the encoder. The side information Y used in the MMSE inverse quantization is generated using the error concealment method in (7) with a uniformly distributed transmission error pattern.

4 Experiments Results

The proposed error resilient depth map coding scheme is implemented based on H.264/AVC (joint model reference software 17.2), and verified by the experiments using multi-view test sequences *Ballet* and *Breakdancers* of which both color video and depth map are provided from Microsoft Research.

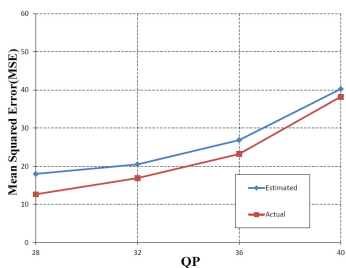


(a) Breakdancer

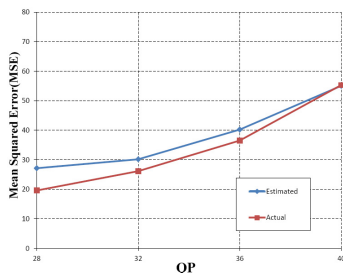


(b) Ballet

Fig. 1. Comparison between the actual and estimated end-to-end distortion. $QP = 30$, packet loss rate = 10%. (a) 50 frames depth map in the 2th view of *Breakdancers*; (b) 50 frames depth map in 2th view of *Ballet*.



(a) Breakdancer



(b) Ballet

Fig. 2. Comparison between the actual and estimated rendered view distortion. Using the 2th view to generate the 3th view. The depth maps used in DIBR are coded as intra mode with $QP=28, 32, 36, 40$.

4.1 End-to-End Distortion Estimation

Firstly, we evaluate the accuracy of the end-to-end distortion estimation model. The depth sequences (50 frames, 1024×768) of the 2th view from two sequences *Ballet* and *Breakdancers* are used. Only the first frame is encoded as I frame, and all the remaining frames are encoded as P frames, $QP = 30$. Each row of macroblocks composes a slice and is transmitted in a separate packet. The packet loss rate is 10%, and we simulate the decoding process 100 times using the error concealment method in (7). Fig. 1 shows the estimated distortion and the actual distortion at every frame. The plots indicated that the estimated distortion is very close to its actual value along the whole sequences.

4.2 Rendered View Distortion Estimation

Both for *Breakdancers* and *Ballet* sequences, we use the color and depth video from the 2th view (50 frames, 1024×768) to generate the rendered image in the

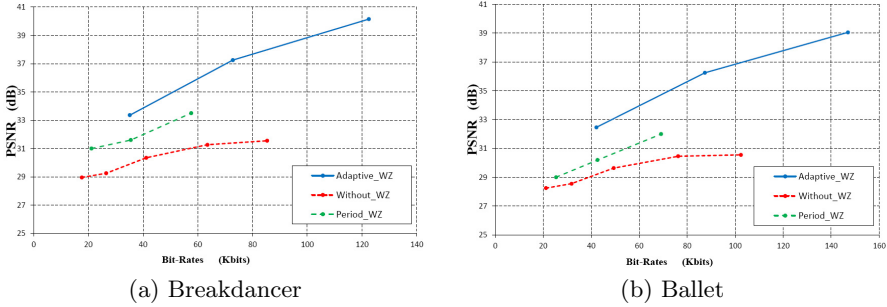


Fig. 3. Comparison results of three coding schemes (using adaptive WZ frames, using periodically WZ frames, without using WZ frames) at packet loss rate is 10%

3th view. The depth maps in the 2th view are all coded as intra mode which will only induce source-coding distortion D_{depth_s} . We estimate $D_{render}(D_{depth_s})$ according to (10) and (11) at encoder. Note that, we use the rendered images generated by the uncompressed depth maps as the ground truth to compute the mean squared error (MSE), not the original images in the 3th view. The results in Fig. 2 shows that the distortion estimation of the rendered view is very high.

4.3 Error Resilient Depth Map Coding

Finally, we evaluate the performance of the proposed error resilient depth map coding scheme using adaptive WZ frame (*Adaptive_WZ*) with our JSC-RDO algorithm. There are other two methods used as baselines for comparison: (1) *Without_WZ*, in this method, only error concealment in (7) and intra refreshment are used. (2) *Period_WZ*, in this method, depth map is periodically coded as WZ frame, and the period is set to 5 in our experiment. Both for *Breakdancers* and *Ballet* sequences, the depth maps in the 2th view (50 frames, 1024×768) are coded and the decoded image will be used to generate the video in the 3th view. The packet loss rate is 10%, and the rendered images generated by the uncompressed depth maps are used as the ground truth. It has to be mentioned that during the WZ coding process, because the resolution of the sequences is not fit for the LDPCA codes, a down-sampling and its inverse process is needed, which will reduce the efficiency at some degree. Fig. 3 shows that the WZ frame can prevent error propagation better than the method in which only error concealment and intra refreshment are used. Obviously, the R-D performance of the adaptively inserted WZ frame outperforms the periodic one significantly and the quality of the rendered view is preserved.

5 Conclusion

We proposed an error resilient depth map coding scheme using adaptive Wyner-Ziv frames, which is based on a joint source-channel R-D optimization (JSC-RDO)

algorithm. The results show that the adaptively inserted WZ frame can outperform the periodic one significantly.

Acknowledgments. The work described in this paper was supported by the National Natural Science Foundation of China (No. 61071184, 60972111, 61036008), Research Funds for the Doctoral Program of Higher Education of China (No. 20100184120009, 20120184110001), Program for Sichuan Provincial Science Fund for Distinguished Young Scholars (No. 2012JQ0029), the Fundamental Research Funds for the Central Universities (Project no. SWJTU09CX032, SWJTU10CX08, SWJTU11ZT08), and Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

References

1. Shimizu, S., Kitahara, M., Kimata, H., Kamikura, K., Yashima, Y.: View Scalable Multiview Video Coding Using 3-D Warping With Depth Map. *IEEE Transactions on Circuits and Systems for Video Technology* 17(11), 1485–1495 (2007)
2. Ndjiki-Nya, P., Koppel, M., Doshkov, D., Lakshman, H., Merkle, P., Muller, K., Wiegand, T.: Depth Image-Based Rendering With Advanced Texture Synthesis for 3-D Video. *IEEE Transactions on Multimedia* 13(3), 453–465 (2011)
3. De Silva, D.V.S.X., Fernando, W.A.C., Worrall, S.T.: 3D Video communication scheme for error prone environments based on motion vector sharing. In: *3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON)*, June 7-9, pp. 1–4 (2010)
4. Liu, Y., Wang, J., Zhang, H.: Depth Image-Based Temporal Error Concealment for 3-D Video Transmission. *IEEE Transactions on Circuits and Systems for Video Technology* 20(4), 600–604 (2010)
5. Yan, B.: A Novel H.264 Based Motion Vector Recovery Method for 3D Video Transmission. *IEEE Transactions on Consumer Electronics* 53(4), 1546–1552 (2007)
6. Zhang, Y., Zhu, C., Yap, K.-H.: A Joint Source-Channel Video Coding Scheme Based on Distributed Source Coding. *IEEE Transactions on Multimedia* 10(8), 1648–1656 (2008)
7. Zhang, Y., Xiong, H., He, Z., Yu, S., Chen, C.W.: An Error Resilient Video Coding Scheme Using Embedded WynerCZiv Description With Decoder Side Non-Stationary Distortion Modeling. *IEEE Transactions on Circuits and Systems for Video Technology* 21(4), 498–512 (2011)
8. Antonio, O., PoLin, L., Dong, T., Cristina, G.: Depth map coding with distortion estimation of rendered view. In: *Proceedings of SPIE-IS and T Electronic Imaging - Visual Information Processing and Communication*, vol. 7543 (2010)
9. Zhang, Y., Gao, W., Lu, Y., Huang, Q., Zhao, D.: Joint Source-Channel Rate-Distortion Optimization for H.264 Video Coding Over Error-Prone Networks. *IEEE Transactions on Multimedia* 9(3), 445–454 (2007)
10. Brites, C., Pereira, F.: Correlation Noise Modeling for Efficient Pixel and Transform Domain WynerCZiv Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology* 18(9), 1177–1190 (2008)
11. Kubasov, D., Nayak, J., Guillemot, C.: Optimal Reconstruction in Wyner-Ziv Video Coding with Multiple Side Information. In: *IEEE 9th Workshop on Multimedia Signal Processing*, October 1-3, pp. 183–186 (2007)

A New Closed Loop Method of Super-Resolution for Multi-view Images

Jing Zhang, Yang Cao, Zhigang Zheng, and Zengfu Wang

University of Science and Technology of China, Hefei, Anhui, P.R. China

`zjwinner@mail.ustc.edu.cn`,

`{forrest,zhengzg,zfwang}@ustc.edu.cn`

Abstract. In this paper, we propose a closed loop method to resolve the multi-view super-resolution problems. Given that the input is one high-resolution view along with its neighboring low-resolution views, our method can give the super-resolution results and obtain a high quality depth map simultaneously. The closed loop method consists of two parts, part I: stereo matching and depth maps fusion and part II: super-resolution. Under the guidance of the depth information, the super-resolution process can be divided into three steps, disparity based pixel mapping, nonlocal construction and final fusion. Once we have the super-resolution results, we can update the disparity maps, and in addition, use the proposed 3D-median filter to update the depth map. We repeat the loop for several times to obtain the high quality super-resolution results and depth map simultaneously. The experimental results show that the proposed method can achieve high quality performance at varies scale factors.

1 Introduction

Single view image super-resolution is a fundamental problem in low-level computer vision research and has been widely studied [1–14]. As shown in [1–3], methods of this problem always belong to the following three categories: interpolation based methods, reconstruction based methods, learning based methods. Interpolation based methods are simple but may blur the edges [4–6]. Reconstruction based methods benefits from different effective constraints or smoothness priors which are consistent with some characters of image [7–10]. Research imposes these constraints or priors to the expected high-resolution target and formulates this super-resolution problem as an optimization problem. The performance of these methods is influenced by the effectiveness of the constraints or priors and the optimization techniques. Learning based methods encode the relationship between high-resolution and low-resolution images in the training set [1, 3, 11–14]. And then it estimates the lost high frequency details in the low-resolution image from the high-resolution images in the training set by using the above relationship. Selection of the training set and proper definition of the relationship are crucial to the performance of these category methods.

These above single view image super-resolution methods may be directly used in stereo and multi-view applications. But if so, they will share a problem that

they do not take advantage of the correspondence of different views. And the performance of the above methods may not meet the expectation. To tackle this problem, Dorea, C and et al. propose an approach to increase the quality of the low-resolution views, using the high frequency information from adjacent high-resolution views which is under the guidance of the available depth information [15]. But, usually the high quality depth map is not available in a direct way, i.e., in mobile stereoscopic camera case. It can be obtained by using some stereo matching algorithms. The problem is that if we only have a mixed resolution views (low-resolution view and its high-resolution stereo counterpart), the stereo matching results often has low quality as stereo algorithms rely on the concurrence of high frequency details in stereo views.

In this paper, to get out of the predicament, we propose a closed loop method to enhance the resolution of multi-view images and obtain a high quality depth map simultaneously. Further more, different from [15], we will tackle the super-resolution problem in a more strict case, which is one high-resolution view along with its neighboring low-resolution views, and without the depth information. Our method consists of two parts shown in Figure 1: part I: stereo matching and depth maps fusion, in this part, we first compute the corresponding disparity maps of different views using stereo matching algorithm. Then, we can get several depth maps of the reference view and fuse them into a more reliable depth map. Part II: image super-resolution. We use the depth map in part I to construct the relationship between the high-resolution view and its neighboring low-resolution view, and then estimate its lost high frequency details from the high-resolution image. Once we obtain the super-resolution results, we can repeat the loop and update the depth map and high-resolution results in these two parts. Finally, we can get the super-resolution results and high quality depth map simultaneously.

The proposed method can be used in 3D video super-resolution and data reduction, multi-view reconstruction, and other video applications [15–17].

2 A New Closed Loop Method of Super-Resolution for Multi-view Images

2.1 Overview

For the super-resolution problem in the mixed-resolution multi-view case (a registered high-resolution image I_1^H along with its several neighboring low-resolution views I_N^L , $N = 2, 3, 4, 5$), our goal is to restore the high-resolution images I_N^H , and obtain a high quality depth map simultaneously. In this paper, we focus on only one kind of multi-view configuration as following: the high-resolution image I_1^H is the leftmost view, and its neighboring low-resolution views I_N^L are the right views in-order. The method for dealing with other configuration with a different order is analogous.

As shown in Figure 1, our closed loop method consists of two parts. The first part aims at fusing several raw depth maps into a more reliable one, and is called stereo matching and depth maps fusion. In this part, those raw depth maps are

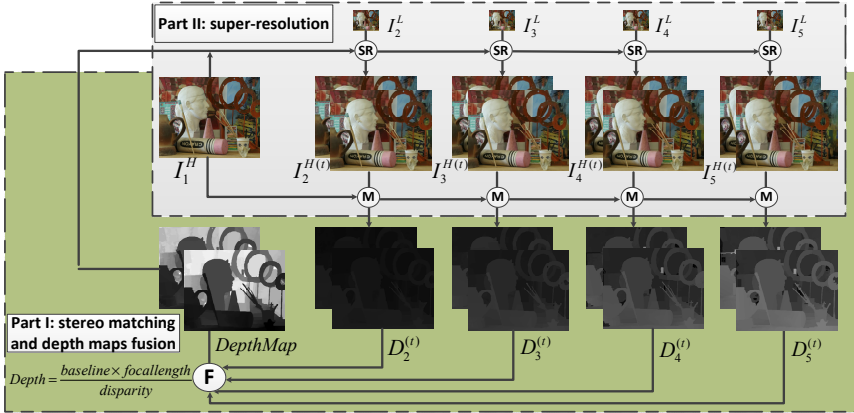


Fig. 1. Flowchart of the closed loop method of super-resolution for multi-view images. Inputs: high-resolution reference view 1 (I_1^H) and its neighboring low-resolution views 2-5 (I_N^L , $N = 2, 3, 4, 5$). Outputs: super-resolution results (I_N^H) of views 2-5 and the depth map.

computed from the disparity maps (D_N) obtained by matching different views I_N^H with the reference view I_1^H . The second part is image super-resolution process. Once we get the fused depth map, we can compute the disparity maps (D'_N) of view 2,3,4,5 relative to view 1. Then under the guidance of these disparity maps, we can estimate the lost high-resolution details of I_N^L from the high-resolution view I_1^H . In addition, we use a nonlocal constraint to utilize the redundancy information in the high-resolution view I_1^H , which is beneficial to super-resolution.

After we get the super-resolution results I_N^H , we use them to update the corresponding disparity maps and obtain a more reliable fused depth map again. We repeat the loop for several times for obtaining good super-resolution results I_N^H and high quality depth map. This closed loop method has two outstanding characteristics. One is that image super-resolution process can benefit from the fused depth map which is gradually becoming more reliable with the repeating. The other one is that stereo matching and depth maps fusion step also benefits from the updated super-resolution results which become to have more reliable high frequency details with the repeating.

2.2 Stereo Matching and Depth Maps Fusion

Under simple binocular camera, stereo matching is consistent with human’s inherent visual perception that we perceive depth based on the disparity of the corresponding point in left eye and right eye [18]. In our multi-view configuration, we can get several disparity maps of the reference view 1 (I_1^H) relative to

other views (I_N^H). Hence, we can compute the depth map according the following formula [18]:

$$Depth_N = \frac{F \times B_N}{D_N} \quad (1)$$

Where, F is the focal length, B_N is the baseline between view 1 and view N , D_N is the disparity map of the reference view 1 relative to view N , and $Depth_N$ is the N^{th} estimation of depth map of view 1.

As we have pointed out before, stereo matching algorithm relies on the concurrence of high frequency details in stereo views. The more distinguish details these two views share, the more reliable the matching results are. But in our super-resolution problem, neighboring views of the high-resolution reference view 1, namely, view 2,3,4,5, are in low resolution. Thus, the disparity map D_N and the corresponding depth map $Depth_N$ may be not reliable. Luckily, we adopt the multi-view configuration described above. So we have four estimations of the depth map of view 1, namely, $Depth_N$, $N = 2, 3, 4, 5$. We can fuse them into a more reliable estimation. Here comes the 3D-median filter that we propose to achieve this goal.

The 3D-median filter is an extension of the traditional 2D-median filter, and its additional dimension is the number of the views in the multi-view configuration. As illustrated in Figure 2, for every position $(i, j) \in A$ on each depth map $Depth_N$, $N = 2, 3, 4, 5$, we firstly select a $r \times r$ patch $P_N(i, j)$ centered on (i, j) , and stack them into a 3D array. A denotes the index set. Then we sort the elements in this 3D array, and chose the median value as the final estimation of depth value on (i, j) . Mathematically,

$$Depth_{fusion}(i, j) = median \{ Depth_N(k, l) | Depth_N(k, l) \in P_N(i, j), N = 2, 3, 4, 5 \} \quad (2)$$

The 3D-median filter can efficiently eliminate the outliers and give a smooth and clear depth map. See Figure 2 for a visual comparison.

2.3 Image Super-Resolution

Once we get the depth map, we can use the formula (1) inversely to compute the corresponding disparity map of view 1 relative to view N . And then, we can interpolate it to get the disparity map D'_N of view N relative to view 1. Once we have the disparity information, we can start the image super-resolution process which is the second part in our closed loop method. The image super-resolution process proposed in this paper consists of three steps: disparity based pixel mapping, nonlocal reconstruction and final fusion.

Disparity Based Pixel Mapping. As we have obtained the disparity map D'_N of view N relative to view 1, a direct way to use it is mapping the pixel value in the high-resolution view I_1^H to the corresponding position of the target super-resolution result I_N^H . However, we know that the disparity value is

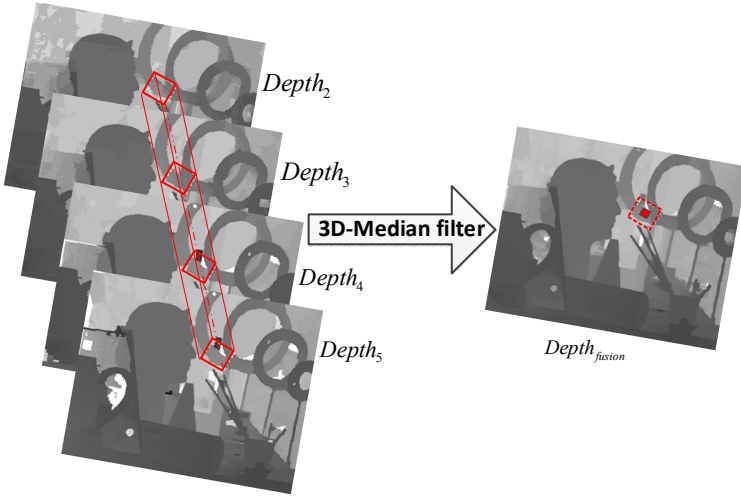


Fig. 2. Illustration of 3D-median filter used for fusing the raw depth maps

not always correct, especially in the occlusion regions and non-overlapping regions of two views. So, firstly we define a confidence value c_{ij} of each disparity to measure whether it is reliable or not. Namely, $c_{ij} = 1$, if the mapping result $I_1^H(i, j + D_N'(i, j))$ exists and the similarity measurement is subjected to $\|TI_N^{H(t)}(i, j) - TI_1^H(i, j + D_N'(i, j))\| \leq \varepsilon$; else $c_{ij} = 0$. T is a vectorization patch extraction operator, and $TI_N^{H(t)}(i, j)$ denotes the vectorization patch centered on (i, j) of $I_N^{H(t)}$. The superscript t records the repeating times of the loop. ε is a similarity threshold. Then, the disparity based pixel mapping can be formulated as:

$$I_{N_mapping}^{H(t+1)}(i, j) = c_{ij} \times I_1^H(i, j + D_N'(i, j)) + (1 - c_{ij}) \times I_N^{H(t)}(i, j) \quad (3)$$

We use linear interpolation to compute the mapping result when the disparity value is not a integer.

Nonlocal Reconstruction. It has been reported that natural image content is likely to repeat itself within some neighborhood [19, 20]. This prior knowledge is very beneficial for resolving super-resolution problems, because it means that we can use much more information to estimate the lost details. In this paper, we use the patch-wise form of nonlocal constraint [20] to help resolve our super-resolution problem. Firstly, for every position (i, j) of the mapping result $I_{N_mapping}^{H(t)}$, we can determine the nonlocal neighborhood of the target position $(i, j + D_N'(i, j))$ of I_1^H . Then, we can select out the patches within this nonlocal neighborhood which are similar with the reference patch $TI_{N_mapping}^{H(t)}(i, j)$

centered on position (i, j) of $I_{N_mapping}^{H(t)}$. Finally, we compute the weight sum of them as the estimation. Mathematically,

$$TI_{N_nonlocal}^{H(t+1)}(i, j) = c_{ij} \times \sum_{(k,l) \in \omega(i, j + D'_N(i, j))} w_{ij,kl} TI_{N_mapping}^{H(t+1)}(k, l) + (1 - c_{ij}) \times TI_{N_mapping}^{H(t+1)}(i, j) \quad (4)$$

Where, $\omega(i, j + D'_N(i, j))$ is the nonlocal neighborhood of the target position $(i, j + D'_N(i, j))$, and $w_{ij,kl}$ is weight computed by measure the similarity between patch $TI_{N_mapping}^{H(t+1)}(k, l)$ and patch $TI_{N_mapping}^{H(t+1)}(i, j)$.

For every position (i, j) , we average all the corresponding pixel values which are involved in those overlapping patches as the final nonlocal reconstruction result $I_{N_nonlocal}^{H(t+1)}(i, j)$. Figure 3 illustrates the nonlocal reconstruction process.

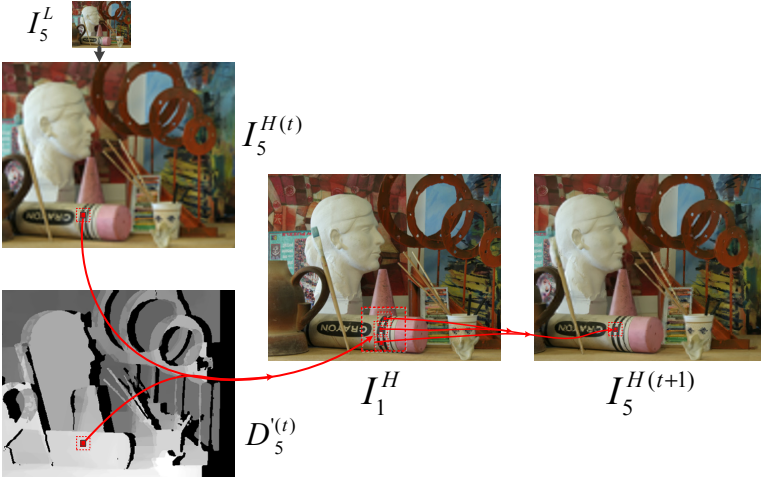


Fig. 3. Illustration of nonlocal reconstruction step in image super-resolution

Final Fusion. While the disparity based mapping result has a little of discontinuities or wrong mapping pixels, the nonlocal reconstruction process can smooth and eliminate them. But the nonlocal reconstruction process may also smooth out some tiny textures (see the face and cup regions in the nonlocal reconstruction result of Figure 3). To compensate for this loss, we fuse the disparity based mapping result and nonlocal reconstruction result and obtain the final super-resolution result. Namely,

$$I_N^{H(t+1)} = \mu_1 \times I_{N_mapping}^{H(t+1)} + \mu_2 \times I_{N_nonlocal}^{H(t+1)} \quad (5)$$

Where μ_1 and μ_2 are two fusion parameters and subjected to $\mu_1 + \mu_2 = 1$ and $0 \leq \mu_1, \mu_2 \leq 1$. In this paper, we set $\mu_1 = \mu_2 = \frac{1}{2}$.

3 Experiments

To test the validity of the proposed closed loop method, we conduct a series of experiments on Middlebury 2005 Datasets¹ [21] with scale factor 2, 4, 8. We give an objective evaluation against bicubic interpolation method and sparse coding method in terms of PSNR and SSIM (Structural Similarity) index [22] and subjective visual comparison. Among those 7 views of each scene in the dataset, we select views 1-5 in our experiments. And the PSNR and SSIM scores are all calculated by averaging the corresponding values of these 4 views (view 2-5, view 1 is the high-resolution reference view). In our experiments, we use the results of sparse coding method as the initial guess to warm up our closed loop method. And we use the stereo matching algorithm proposed in [23] to obtain the disparity map.

3.1 Super-Resolution Results

Table 1 and Table 2 summarize PSNR and SSIM scores of the three methods with different scale factors. It is clear that our approach consistently outperforms bicubic interpolation method and sparse coding method across all the scale factors and images. Our approach achieves average gains of 2.47, 2.99 and 1.42 dB over bicubic interpolation method for scale factor 2, 4, 8 respectively. And compared with sparse coding method, average PSNR gains are 0.83, 2.05 and 0.79 dB for scale factor 2, 4, 8 respectively. The SSIM scores of our method also reflect significant gains over the bicubic interpolation method and sparse coding method.

The subjective visual comparisons also confirm the superiority of our method. Figure 4 shows the super-resolution results of Art (view 2) in the dataset with a scale factor 4. And Figure 5 shows the super-resolution results of Reindeer (view 3) in the dataset with a scale factor 8. It can be seen that the results of our method are much more visually pleasing and have more and finer high-frequency detail than the other two methods. We recommend viewing these figures on a screen.



Fig. 4. Super-resolution results of Art (view 2) with a scale factor 4. (a), the low-resolution input. (b), result of bicubic interpolation method. (c), result of sparse coding method. (d), result of proposed method. (e) the groundtruth.

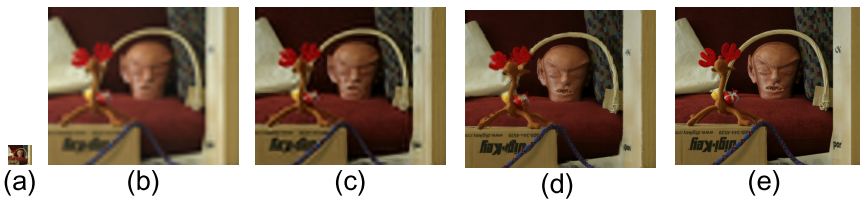
¹ <http://vision.middlebury.edu/stereo/data/>

Table 1. PSNR scores (dB) of super-resolution results obtained by different methods on Middlebury 2005 datasets

Scale factor	Method	Art	Books	Dolls	Laundry	Moebius	Reindeer
2	Bicubic	32.65	29.13	31.73	30.55	33.16	32.76
	Sparse Coding	34.56	30.62	33.50	32.39	34.51	34.27
	Proposed	35.24	31.44	34.67	32.76	35.51	35.18
4	Bicubic	27.30	24.85	26.61	25.28	28.58	28.50
	Sparse Coding	28.47	25.78	27.68	25.87	29.26	29.69
	Proposed	30.55	28.07	30.60	27.08	31.20	31.53
8	Bicubic	23.40	21.38	22.73	21.76	25.01	24.99
	Sparse Coding	23.99	21.99	23.39	22.11	25.49	26.12
	Proposed	24.15	22.91	24.78	22.33	26.38	27.25

Table 2. SSIM scores of super-resolution results obtained by different methods on Middlebury 2005 datasets

Scale factor	Method	Art	Books	Dolls	Laundry	Moebius	Reindeer
2	Bicubic	0.9333	0.8739	0.9237	0.9005	0.9170	0.9196
	Sparse Coding	0.9500	0.8976	0.9427	0.9185	0.9357	0.9313
	Proposed	0.9574	0.9216	0.9575	0.9287	0.9527	0.9439
4	Bicubic	0.7917	0.7166	0.7676	0.7473	0.7914	0.8202
	Sparse Coding	0.8208	0.7398	0.7982	0.7743	0.8139	0.8387
	Proposed	0.9100	0.8835	0.9162	0.8659	0.9082	0.9053
8	Bicubic	0.6203	0.5663	0.5811	0.5898	0.6469	0.6989
	Sparse Coding	0.6365	0.5775	0.6036	0.5998	0.6593	0.7218
	Proposed	0.733	0.7546	0.7746	0.7282	0.7917	0.8342

**Fig. 5.** Super-resolution results of Reindeer (view 3) with a scale factor 8. (a), the low-resolution input. (b), result of bicubic interpolation method. (c), result of sparse coding method. (d), result of proposed method. (e) the groundtruth.

3.2 High Quality Depth Map

As we the flowchart in Figure 1 indicates, our closed loop method can also obtain depth map simultaneously. To examine the quality of the depth results, we use the relative error of depth map (REoD) to measure how good a depth map is when compared with the groundtruth. REoD is calculated as:

$$REoD = \frac{1}{|A|} \sum_{(i,j) \in A} \frac{|Depth(i,j) - Depth_{groundtruth}(i,j)|}{|Depth_{groundtruth}(i,j)|} \quad (6)$$

Since we use the results of sparse coding method as the initial guess, thus we have the corresponding disparity maps by using stereo matching algorithm. Then, we can get four depth maps according to formula (1). We calculate the corresponding REoDs of these four initial depth maps and average them. The results are shown in Table 3 indicated by “Initial”. Generally, after 1-3 times’ repeating, the closed loop method gives pleasing super-resolution results and depth maps. The REoDs of these depth maps are shown in Table 3 indicated by “Proposed”. In addition, we match the reference high-resolution view 1 with the high-resolution groundtruth views 2-5 to get the corresponding depth maps. And we calculate the REoDs of these depth maps. They are shown in Table 3 indicated by “MoG”(Matching results of Groundtruth). These results can be used to compare with results of our method. Further more, we can know the goodness of the super-resolution results of our method indirectly from such a comparison.

REoD values in Table 3 shows that our method significantly improves the quality of the initial depth maps. And it gives comparable results to “MoG” for scale factor 2 and 4. Although the values of our method for scale factor 8 are still a little higher, but they are much lower than the values of the initial maps. Generally, these comparison results convince the effectiveness of our method on the aspect of obtaining a high quality depth map. Figure 6 and Figure 7 shows the depth map results along with the above super-resolution experiments shown in Figure 4 and Figure 5. It can be seen that our method improves the initial depth maps and gives comparable results to “MoG” and the groundtruth.

Table 3. REoD values of the depth map results obtained by different methods on Middlebury 2005 datasets

Scale factor	Method	Art	Books	Dolls	Laundry	Moebius	Reindeer
2	Initial	0.0899	0.1772	0.0443	0.0856	0.0710	0.0475
	Proposed	0.0321	0.0285	0.0193	0.0375	0.0351	0.0314
4	Initial	0.1106	0.1577	0.0909	0.1109	0.0791	0.0627
	Proposed	0.0454	0.0681	0.0314	0.0523	0.0392	0.0518
8	Initial	0.2878	0.2307	0.1565	0.2540	0.1764	0.1129
	Proposed	0.0874	0.0658	0.0482	0.1003	0.1011	0.0638
1	MoG	0.0305	0.0247	0.0184	0.0440	0.0312	0.0291

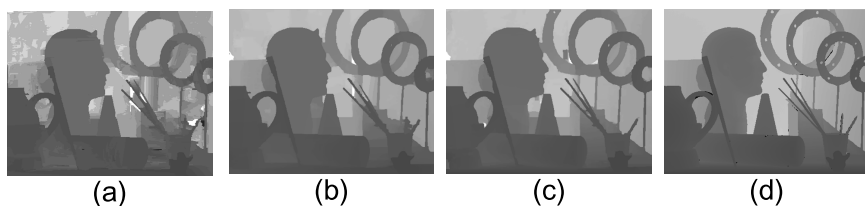


Fig. 6. Depth map results of Art (scale factor: 4). (a), Initial depth map. (b), depth map result of proposed method. (c)depth map result of MOG. (d), the groundtruth.

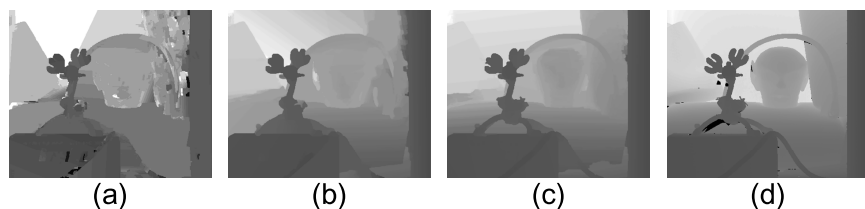


Fig. 7. Depth map results of Reindeer (scale factor: 8). (a), Initial depth map. (b), depth map result of proposed method. (c)depth map result of MOG. (d), the groundtruth.

4 Conclusion

In this paper, to tackle the multi-view super-resolution problem, we propose a closed loop method. This novel method shows its superiority of obtaining impressive super-resolution results and high quality depth map simultaneously. The major characteristic of this method is updating the depth map and super-resolution results alternately, and each of them benefits from the improvement of the other. Series of Experiments confirms the effectiveness of this method. The proposed method may be used in the 3D video super-resolution problem for further study. The future work may also concentrate on hardware implement of the special multi-view configuration which is a high-resolution camera along with several low-resolution cameras, and the algorithm optimization for constructing the high-accuracy 3D structure of the scene.

Acknowledgements. This paper is supported by the Fundamental Research Funds for the Central Universities of China (No. WK2100100009), NSFC (No.61175033), NSFY (No.BJ2100100018) and STP (No.11010202192) of Anhui.

References

1. Sun, J., Xu, Z., Shum, H.-Y.: Image super-resolution using gradient profile prior. In: CVPR (2008)
2. He, H., Siu, W.C.: Single image super-resolution using gaussian processing regression. In: CVPR (2011)

3. Tai, Y.W., Liu, S., et al.: Super resolution using edge prior and single image detail synthesis. In: CVPR (2010)
4. Allebach, J., Wong, P.: Edge-directed interpolation. In: ICIP (1996)
5. Caselles, V., Morel, J.M., Sbert, C.: An axiomatic approach to image interpolation. *IEEE Trans. on Image Processing* 7(3), 376–386 (1998)
6. Li, X., Orchard, M.: New edge-directed interpolation. In: ICIP (2000)
7. Aly, H.A., Dubois, E.: Image up-sampling using total-variation regularization with a new observation model. *IEEE Trans. on Image Processing* 14(10), 1647–1659 (2005)
8. Dai, S., Han, M., et al.: Soft edge smoothness prior for alpha channel super resolution. In: CVPR (2007)
9. Shan, Q., Li, Z., Jia, J., Tang, C.K.: Fast image/video upsampling. *ACM Trans. On Graphics, SIGGRAPH ASIA* (2008)
10. Protter, M., Elad, M., Takeda, H., Milanfar, P.: Generalizing the nonlocal-means to super-resolution reconstruction. *IEEE Trans. on Image Processing* 18(1), 36–51 (2009)
11. Baker, S., Kanade, T.: Limits on super-resolution and how to break them. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 24(9), 1167–1183 (2002)
12. Chang, H., Yeung, D.-Y., Xiong, Y.: Super-resolution through neighbor embedding. In: CVPR (2004)
13. Wang, Q., Tang, X., Shum, H.Y.: Patch based blind image super resolution. In: ICCV (2005)
14. Yang, J., Wright, J., Huang, T., Ma, Y.: Image super-resolution via sparse representation. *IEEE Trans. on Image Processing* 19(11), 2861–2873 (2010)
15. Garcia, D., Dorea, C., de Queiroz, R.: Super resolution for multiview images using depth information. *IEEE Trans. on Circuits and Systems for Video Technology* 22(9), 1249–1256 (2012)
16. Wang, M., Hua, X.-S., et al.: Beyond distance measurement: Constructing neighborhood similarity for video annotation. *IEEE Trans. on Multimedia* 11(3), 465–476 (2009)
17. Wang, M., Hong, R., et al.: Event driven web video summarization by tag localization and key-shot identification. *IEEE Trans. on Multimedia* 14(4), 975–985 (2012)
18. Szeliski, R.: *Computer vision: algorithms and applications*. Springer (2011)
19. Buades, A., Coll, B., Morel, J.M.: A non-local algorithm for image denoising. In: CVPR (2005)
20. Buades, A., Morel, B.C., Morel, J.M.: A review of image denoising algorithms with a new one. *Multiscale Model. Simul.* 4, 490–530 (2005)
21. Scharstein, D., Pal, C.: Learning conditional random fields for stereo. In: CVPR (2007)
22. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. *IEEE Trans. on Image Processing* 13(4), 600–612 (2004)
23. Wang, Z.F., Zheng, Z.G.: A region based stereo matching algorithm using cooperative optimization. In: CVPR (2008)

Fast Coding Unit Decision Algorithm for Compressing Depth Maps in HEVC

Yung-Hsiang Chiu¹, Kuo-Liang Chung¹, Wei-Ning Yang²,
Yong-Huai Huang^{3,*}, and Chih-Ming Lin¹

¹ Department of Computer Science and Information Engineering,
National Taiwan University of Science and Technology,
No. 43, Section 4, Keelung Road, Taipei, Taiwan 10672, R. O. C.

² Department of Information Management,
National Taiwan University of Science and Technology,
No. 43, Section 4, Keelung Road, Taipei, Taiwan 10672, R. O. C.

³ Institute of Computer and Communication Engineering,
Jinwen University of Science and Technology,
No. 99, Anzhong Road, Xindian District, New Taipei City, Taiwan 23154, R. O. C.
yonghuai@ms28.hinet.net

Abstract. The stereoscopic video system creates a 3-D scene using the color map and the virtual view which is synthesized by the color and depth maps. Compressing 3-D video sequences is a crucial research issue. Based on Zhao *et al.*'s depth no-synthesis-error model, this paper presents a fast coding unit decision algorithm for compressing depth maps in HEVC (High Efficiency Video Coding). The proposed coding unit decision algorithm determines, as early as possible, the minimum coding unit size required in the quadtree structure of HEVC while preserving an acceptable quality. Experimental results demonstrate that the proposed coding unit decision algorithm computationally outperforms the one used in HEVC while incurring negligible degradation on both quality and bitrate performance.

Keywords: Coding unit decision, Depth-image-based rendering (DIBR), Depth no-synthesis-error (D-NOSE) model, HEVC, Quadtree structure, Time performance.

1 Introduction

3-D TV systems [1] have received growing attention in consumer electronics market. The systems aim to provide viewers with the perception of stereoscopic scene which is achieved by the left and right virtual views depicted to human eyes. There are three kinds of 3-D video formats, namely, color plus depth map [2], stereo [3], and multi-view formats [4]. Compared to the stereo and multi-view videos, the color plus depth map format requires lower storage and transmission bandwidth and hence is usually considered for 3-D TV broadcasting. In this

* Corresponding author.

research, we focus on the color plus depth map format. Each time we have two maps, one color map and one depth map. According to the depth-image-based rendering (DIBR) technique [2], stereo pairs of the left and right virtual views, created by synthesizing the color map and the corresponding depth map, are used to form the 3-D video.

In 3-D TV broadcasting, since transmitting on the internet the stream of color plus depth video may require large amount of resources, the issue of compressing a color plus depth map 3-D video has captured the interests of researchers in recent years. Previously, in H.264/AVC, several articles have been published to tackle the compression issues of color map [5–7], depth map [8, 9], and both maps.

High Efficiency Video Coding (HEVC) is the next generation compression standard for video coding. Compared to H.264/AVC, HEVC achieves superior bitrate and quality performance at the expense of higher computational time complexity. In HEVC, the coding unit decision process, which involves determining the partition size of coding units and selecting prediction mode for each coding unit, involves large computational cost and plays a crucial role. Since traditional coding unit decision process involves exhaustive search which incurs large computational cost, most researches focus on speeding up the coding unit decision process with little degradation on the rate-distortion (RD) cost. However, the existing fast coding unit decision algorithms [10–12] for color images focus on determining the partition size of coding unit in the quadtree structure, where each coding unit is divided into four subsequent coding units, initially designed for gray images, and there exist no coding unit decision algorithms for color plus depth maps.

In this paper, we propose a novel fast coding unit decision algorithm for compressing depth maps in HEVC. The proposed algorithm speeds up the coding unit decision process by determining, based on the D-NOSE model [13], the minimum coding unit size required in the quadtree structure of HEVC as early as possible while preserving an acceptable quality. The coding unit decision process stops partitioning the coding units once the reconstructed depth values fall in the allowable range determined by the D-NOSE model to speed-up the encoding process. Since the reconstructed depth values are within the allowable range, no synthesis errors will be observed, resulting in little degradation on the quality of 3D scenes. Experimental results demonstrate that the proposed algorithm outperforms computationally the one used in HEVC while incurring negligible degradation on both quality and bitrate performance.

2 View Synthesis for 3-D TV

The depth-image-based rendering (DIBR) technique [2] is used for view synthesis in 3-D video system. DIBR consists of two key steps, namely, warping and hole-filling. Warping step maps the original color map, which is viewed as the original view, to the virtual view based on the associated depth map. Since warping is not a one-to-one mapping, hole-filling is used to polish the virtual view. Since

our research focuses on coding unit decision process in HEVC for compressing depth maps, we only survey the warping technique in this section.

The warping procedure transforms a pixel with the coordinate $m \equiv (x, y)^T$ in the original view to a pixel $m' \equiv (x', y')^T$ in the right virtual view. For simplicity, we only consider the horizontal disparity between the original and the virtual views and the horizontal disparity is denoted by $x' - x$. Assuming that the world coordinate system equals the camera coordinate system of the camera, the perspective projection equation projects the 3-D space point $M = (X, Y, Z)^T$ to the 2-D image point $m = (x, y)^T$ with depth value $z = (Z)$ in the original view; it yields

$$z \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = AM \quad (1)$$

where

$$A = \begin{bmatrix} f & \tau & o_x \\ 0 & \eta f & o_y \\ 0 & 0 & 1 \end{bmatrix} \quad (2)$$

denotes the intrinsic camera parameter matrix with f denoting the focal length, $(o_x, o_y)^T$ denoting the principal-point position, and parameters η and τ modeling the aspect ratio and skew of pixels. In practice, when employing modern digital cameras, it can be safely assumed that pixels are square ($\eta = 1$) and non-skewed ($\tau = 0$).

And the perspective projection equation which projects the 3-D space point $M = (X, Y, Z)^T$ to the 2-D image point $m' = (x', y')^T$ with depth value z' in the virtual view can be expressed as

$$z' \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = ARM + At \quad (3)$$

where $R = I$ denotes the rotation matrix, i.e. the identity matrix under the horizontal disparity assumption, and $t = [\ell, 0, 0]^T$ represents the translation vector with ℓ being the baseline length between the original and virtual views. If the original view lies to the left with respect to the virtual view, ℓ assumes negative values.

Rearranging Eq. (1) and substituting into into Eq. (3) gives the classical affine disparity equation, which defines the depth-dependent relation between corresponding points $(x, y)^T$ and $(x', y')^T$ in the original and virtual views of the same 3D scene,

$$z' \begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = zI \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} + At, \quad (4)$$

implying that the horizontal disparity depends on the depth value and the pixel position in the original view. Further, the affine disparity equation corresponding the the case of generating a stereo pair of views can be written as

$$\begin{pmatrix} x' - x \\ y' - y \\ z' - z \end{pmatrix} = \begin{pmatrix} f\ell/z \\ 0 \\ 0 \end{pmatrix}. \quad (5)$$

We thus have

$$\begin{pmatrix} x' \\ y' \\ 0 \end{pmatrix} = \begin{pmatrix} x + d_x \\ y \\ 0 \end{pmatrix}, \quad (6)$$

where $d_x = f\ell/z$ denotes the associated horizontal disparity.

For saving the storage, the depth value z is usually converted to a quantized depth value z_q by the quantization function used in MPEG-3DV [14]

$$z_q = Q(z) = \left\lfloor 255 \times \frac{Z_{near}}{z} \times \frac{Z_{far} - z}{Z_{far} - Z_{near}} + 0.5 \right\rfloor \quad (7)$$

where Z_{near} and Z_{far} are the nearest and farthest real depth values in the original view. Values 0 and 255 of the quantized depth value represent, respectively, the farthest and nearest distance from the camera. For synthesizing the virtual view, the quantized depth value z_q is dequantized by

$$z = Q^{-1}(z_q) = \frac{1}{\frac{z_q}{255} \left(\frac{1}{Z_{near}} - \frac{1}{Z_{far}} \right) + \frac{1}{Z_{far}}}, \quad (8)$$

which is used for calculating the horizontal disparity to generate the virtual view in the warping process.

3 Fast Coding Unit Decision Algorithm Based on D-NOSE Model

When generating the virtual view using the horizontal disparity, the horizontal disparity is rounded to the nearest integer value; that is, different depth values may result in the same horizontal disparity, indicating that, for each pixel, there exists a depth value range which generates the common pixel in the virtual view. Zhao *et al.* [13] proposed the D-NOSE model which determines the allowable range of depth values for smoothing to generate the virtual view without any synthesis errors. In this section, we first introduce the D-NOSE model for determining the allowable range of depth values, and then describe the coding unit decision process in HEVC; finally we present how to use the allowable range for speed up the coding unit decision process.

3.1 Depth No-Synthesis-Error Model

In the D-NOSE model, for depth value z_q there exist an allowable range $[\ell(z_q), u(z_q)]$ within which varying depth value does not result in any synthesis errors. When performing warping, the quantized depth value z_q is first dequantized to z by the dequantization function in Eq. (8). The dequantized depth

value z is then used for calculating the horizontal disparity $d_x = f\ell/z$. The λ -rounding ($0 < \lambda \leq 1$) with integer-precision equation for horizontal disparity d_x can be expressed as

$$R(d_x) = \lceil d_x - \lambda \rceil = \left\lceil \frac{f\ell}{Q^{-1}(z_q)} - \lambda \right\rceil, \quad (9)$$

where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x .

Since the quantized depth value z_q and the rounded disparity $R(d_x)$ are not one-to-one transformation, there may exist different quantized depth values corresponding to the common rounded disparity; that is, given a specific quantized depth value z_q , there exists a range $Rg(z_q)$ for the quantized depth value corresponding to the common rounded disparity or

$$Rg(z_q) = \left\{ v \mid \left\lceil \frac{f\ell}{Q^{-1}(v)} - \lambda \right\rceil = \left\lceil \frac{f\ell}{Q^{-1}(z_q)} - \lambda \right\rceil \right\}. \quad (10)$$

Since $R(d_x)$ is a monotonically non-increasing function, for quantized depth value z_q , there exists a unique range $[\ell(z_q), u(z_q)] = [\min(Rg(z_q)), \max(Rg(z_q))]$ such that no synthesis error will be observed in view rendering if the quantized depth value falls in this range.

3.2 The Coding Unit Decision Process in HEVC

When compressing 3D video sequences using the HEVC, coding units are the basic units to be encoded for the depth map of each frame. The size of a coding unit can vary from 64×64 to 8×8 . Using smaller coding units usually results in less distortion but higher bitrate.

In HEVC, the RD cost measures the tradeoff between the bitrate and the distortion and often serves as the performance measure when determining the coding unit size. The coding unit decision process determines the optimal coding unit size and corresponding prediction mode implemented in HEVC to minimize the RD cost. The HEVC first partitions depth map of each frame into a set of coding units of size 64×64 . The 64×64 coding unit structure is represented by a quadtree. Denote by $C_{2N_0 \times 2N_0}^0$ with $N_0 = 32$ the 64×64 coding unit which can be partitioned into four 32×32 coding units $C_{2N_1 \times 2N_1}^0$, $C_{2N_1 \times 2N_1}^1$, $C_{2N_1 \times 2N_1}^2$, and $C_{2N_1 \times 2N_1}^3$ with $N_1 = 16$. Furthermore, the 64×64 coding unit can be partitioned into sixteen 16×16 coding units $C_{2N_2 \times 2N_2}^0$, $C_{2N_2 \times 2N_2}^1$, \dots , $C_{2N_2 \times 2N_2}^{15}$ with $N_2 = 8$ by partitioning each 32×32 coding unit into four 16×16 coding units. Finally, the 64×64 coding unit are partitioned into 64 8×8 coding units $C_{2N_3 \times 2N_3}^0$, $C_{2N_3 \times 2N_3}^1$, \dots , $C_{2N_3 \times 2N_3}^{63}$ with $N_3 = 4$.

For each coding unit in the quadtree, three prediction modes, called intra, inter, and skip, are available for encoding depth values. There exist one, two, and four partitioning patterns for skip, intra, and inter prediction modes, respectively. Exploiting the spatial redundancy of depth maps, intra prediction mode predicts the depth values in each coding unit by referring the depth values from the neighboring encoded coding units. When performing intra prediction

for the coding units $C_{2N_j \times 2N_j}^i$, $i = 0, 1, \dots, 2^{2j} - 1$, two partitioning patterns $P_{\text{intra}, 2N_j \times 2N_j}^i$ and $P_{\text{intra}, N_j \times N_j}^i$ are available. Instead of using the spatial redundancy, inter prediction mode exploits the temporal redundancy of depth maps and predicts the depth values for each coding unit based on the previous encoded depth maps. When performing inter prediction for the coding units $C_{2N_j \times 2N_j}^i$, $i = 0, 1, \dots, 2^{2j} - 1$, four partitioning patterns $P_{\text{inter}, 2N_j \times 2N_j}^i$, $P_{\text{inter}, N_j \times 2N_j}^i$, $P_{\text{inter}, 2N_j \times N_j}^i$, and $P_{\text{inter}, N_j \times N_j}^i$ are available. For the coding units in the still background region, no depth values are recorded and the skip prediction mode $P_{\text{skip}, 2N_j \times 2N_j}^i$ for $i = 0, 1, \dots, 2^{2j} - 1$ is often used.

For each coding unit $C_{2N_j \times 2N_j}^i$, denote by $S_{2N_j \times 2N_j}^i = \{P_{\text{skip}, 2N_j \times 2N_j}^i, P_{\text{intra}, 2N_j \times 2N_j}^i, P_{\text{intra}, N_j \times N_j}^i, P_{\text{inter}, 2N_j \times 2N_j}^i, P_{\text{inter}, N_j \times 2N_j}^i, P_{\text{inter}, 2N_j \times N_j}^i, P_{\text{inter}, N_j \times N_j}^i\}$, the set of available prediction patterns. The prediction pattern for encoding $C_{2N_j \times 2N_j}^i$ is determined to minimize the RD cost and the RD cost corresponding to coding unit $C_{2N_j \times 2N_j}^i$ can be calculated by

$$RD(C_{2N_j \times 2N_j}^i) = \min_{P \in S_{2N_j \times 2N_j}^i} \{D(P) + \lambda R(P)\} \quad (11)$$

where P represents the prediction pattern, $D(P)$ and $R(P)$ denote, respectively, the corresponding distortion and bitrate, and λ is the Lagrange multiplier.

Once the RD costs corresponding to all possible coding units are determined, the quadtree corresponding to the 64×64 coding unit, which specifies how the 64×64 coding unit is partitioned, can be determined as follows.

Consider the coding unit $C_{2N_j \times 2N_j}^i$ with four child coding units $C_{2N_{j+1} \times 2N_{j+1}}^{4i}$, $C_{2N_{j+1} \times 2N_{j+1}}^{4i+1}$, $C_{2N_{j+1} \times 2N_{j+1}}^{4i+2}$, and $C_{2N_{j+1} \times 2N_{j+1}}^{4i+3}$. If the sum of the RD costs of four child coding units is smaller than that of coding unit $C_{2N_j \times 2N_j}^i$, then the coding unit $C_{2N_j \times 2N_j}^i$ is replaced by four child coding units with RD cost equal to the sum of four RD costs; otherwise, retain the coding unit $C_{2N_j \times 2N_j}^i$ and discard the four child coding units. Thus the quadtree corresponding to the 64×64 coding unit can be determined by iterating the above process, starting from coding units $C_{2N_3 \times 2N_3}^i$ for $i = 0, 1, \dots, 63$.

The above coding unit decision process involves calculating the RD cost, determining the prediction pattern for each partitioned coding unit of size from 64×64 to 8×8 , resulting in high computational complexity. In next subsection, based on the D-NOSE mode, we present a new fast coding unit decision algorithm which early stops partitioning the coding units to speed up the encoding process with little error.

3.3 Fast Coding Unit Decision Algorithm for Compression Depth Maps

The coding unit decision process in HEVC partitions each coding unit of size from 64×64 to 8×8 and evaluates the corresponding RD costs to determine the optimal size. However, since the stereoscopic video system purposes to construct the virtual view in the decoder site, instead of examining all possible coding

unit size in the quadtree structure to minimize the sum of the RD costs, we aim to determine, as early as possible, the minimal coding unit size required in quadtree structure while preserving the quality of virtual view and speeding up the coding unit decision process.

In the coding unit decision process, each partitioned coding unit has a best prediction pattern which minimizes the RD cost and the reconstructed depth values from the best prediction pattern do not result in any synthesis error if they all fall in the allowable range determined by the D-NOSE model. Based on the D-NOSE mode, for the depth value z_q , it exists an allowable range $[\ell(z_q), u(z_q)] = [\min(Rg(z_q)), \max(Rg(z_q))]$ obtained by Eq. (10) such that no synthesis errors will be observed in virtual view. For convenience, if the coding unit $C_{2N_j \times 2N_j}^i$, $i = 0, 1, 2, \dots, 2^{2j-1}$ and $j = 0, 1, 2, 3$, whose reconstructed depth values are all fallen in the range $[\ell(z_q), u(z_q)]$, we call that $C_{2N_j \times 2N_j}^i$ satisfies the D-NOSE condition.

Once $C_{2N_j \times 2N_j}^i$ satisfies the D-NOSE condition, the partition process for $C_{2N_j \times 2N_j}^i$ can be stopped to improve the encoding time performance. For each of 64×64 , 32×32 , 16×16 , and 8×8 coding units, we perform some experiments to calculate the ratio of coding units satisfying the D-NOSE condition. Four multi-view sequences of color plus depth map format, including two HD sequences, Kendo and Balloons, and two full HD sequences, Poznan Street and Dancer, were used in our experiments. Four depth map sequences are constructed from the third view of Kendo sequence, the first view of Balloons sequence, the fourth view of Poznan Street sequence, and the second view of Dancer sequence. After compressing the four depth map sequences by HEVC for QP = 24, the ratios of different sized coding units which satisfy the D-NOSE condition are shown in Table 1.

Table 1. Ratios of different sized coding units satisfying the D-NOSE condition

Sequence Name	Kendo	Balloons	PoznanStreet	Dancer	Average
64×64	39.21%	32.75%	4.55%	75.62%	38.03%
32×32	61.69%	49.83%	19.02%	87.47%	54.50%
16×16	78.57%	68.54%	48.56%	94.08%	72.44%
8×8	88.75%	82.64%	73.66%	97.15%	85.55%

Table 1 reveals that on average 38.03%, 54.50%, 72.44%, and 85.55% of 64×64 , 32×32 , 16×16 , and 8×8 coding units, respectively, satisfy the D-NOSE condition, and it indicates that our proposed D-NODE-based early coding unit decision approach has a significant encoding time reduction benefit.

The proposed fast coding unit decision algorithm follows the conventional quadtree-based partition process but the proposed algorithm early terminates the partition by checking the D-NOSE condition resulting in smaller quadtree than that in HEVC. The detailed proposed coding unit decision algorithm for each 64×64 coding unit $C_{2N_0 \times 2N_0}^0$ is described as follows.

1. Coding unit partition stage:

Initialize i and j by setting $i = 0$ and $j = 0$. Perform (a), (b), and (c) to partition the coding unit $C_{2N_0 \times 2N_0}^0$.

- (a) For the coding unit $C_{2N_j \times 2N_j}^i$, find the prediction pattern P , $P \in S_{2N_j \times 2N_j}^i$ and $S_{2N_j \times 2N_j}^i$ is the set of available prediction patterns defined in Eq. (11), such that the RD cost (see Eq. (11)) is minimized. If $j = 3$, i.e. the coding unit size is 8×8 , stop the partition process of $C_{2N_j \times 2N_j}^i$; otherwise, perform (b).
- (b) Reconstruct $C_{2N_j \times 2N_j}^i$ according to the prediction pattern P . The reconstructed depth values \hat{z}_q 's check whether $C_{2N_j \times 2N_j}^i$ satisfies

$$\hat{z}_q \in [\min(Rg(z_q)), \max(Rg(z_q))], \forall z_q, \hat{z}_q \text{ in } C_{2N_j \times 2N_j}^i$$

If the condition is held, stop the partition process of $C_{2N_j \times 2N_j}^i$; otherwise, perform (c).

- (c) Partition $C_{2N_j \times 2N_j}^i$ into four child coding units $C_{2N_{j+1} \times 2N_{j+1}}^{4i}$, $C_{2N_{j+1} \times 2N_{j+1}}^{4i+1}$, $C_{2N_{j+1} \times 2N_{j+1}}^{4i+2}$, and $C_{2N_{j+1} \times 2N_{j+1}}^{4i+3}$. For each child coding unit, perform (a) to determine the prediction pattern and check whether the D-NOSE condition is held or not.

2. Coding unit decision stage:

Starting from the smallest coding units of size $2N_{j+1} \times 2N_{j+1}$ in the quadtree structure constructed by the coding unit partition stage, if the sum of the RD costs of four $2N_{j+1} \times 2N_{j+1}$ coding units $C_{2N_{j+1} \times 2N_{j+1}}^{4i}$, $C_{2N_{j+1} \times 2N_{j+1}}^{4i+1}$, $C_{2N_{j+1} \times 2N_{j+1}}^{4i+2}$, and $C_{2N_{j+1} \times 2N_{j+1}}^{4i+3}$, is smaller than that of their parent coding unit $C_{2N_j \times 2N_j}^i$, then the coding unit $C_{2N_j \times 2N_j}^i$ is replaced by its four child coding units; otherwise, retain the coding unit $C_{2N_j \times 2N_j}^i$ and discard its four child coding units. The above process is performed iteratively until the 64×64 coding unit is reached.

4 Experimental Results

We expect the proposed algorithm can substantially reduce the encoding time with slight degradation in the video quality and bitrate. To verify the effectiveness of the proposed algorithm, we compare the performance of the proposed algorithm with the conventional HEVC in terms of video quality, bitrate, and encoding time. In accordance with the experiments for testing the DNOSE condition mentioned in section 3.3, four depth map sequences extracted from four multi-view sequences in color plus depth map format, namely Kendo, Balloons, Poznan Street and Dancer, are used to evaluate the performance of the proposed algorithm. In addition, four color sequences extracted from the fourth view of Kendo sequence, the second view of Balloons sequence, the fifth view of Poznan Street sequence, and the third view of Dancer sequence are used as the ground truth of the synthesized virtual view sequences.

All experiments were performed on the computer with Intel i7-3770 CPU 3.4 GHz and 16GB RAM. The program developing environment is Visual Studio C++ 2008 implemented on platform HM 4.0 with Microsoft Windows 7 operating system. To clearly report the experimental results, we first define the performance measures as follows. The quality degradation can be measured by the decrease in $PSNR$ when using the proposed algorithm and is defined as $\Delta PSNR = PSNR_{proposed} - PSNR_{HEVC}$ where $PSNR_{HEVC}$ denotes the video quality by using the conventional HEVC and $PSNR_{proposed}$ the video quality by using the proposed algorithm. The percentage of time saving when using the proposed algorithm over the conventional HEVC is defined as $\Delta R_T = \frac{T_{HEVC} - T_{proposed}}{T_{HEVC}} \times 100\%$ where T_{HEVC} denotes the total encoding time required in the conventional HEVC and $T_{proposed}$ the total encoding time required in the proposed algorithm. The increase percentage in bitrate by the proposed algorithm is defined as $\Delta R_B = \frac{B_{proposed} - B_{HEVC}}{B_{HEVC}} \times 100\%$ where B_{HEVC} denotes the total bitrate needed in the conventional HEVC and $B_{proposed}$ the total bitrate needed in the proposed algorithm.

The empirical results of the three performance measures, $\Delta PSNR$, ΔR_T and ΔR_B for different values of QP are listed in Table 2. The average performance measures over different test sequences and different values of QP are also tabulated. From Table 2, we found that the proposed algorithm can achieve significant reduction in the encoding-time with little degradation in RSNR and bitrate when compared with the conventional HEVC. On average the proposed algorithm can achieve 21.19% to 58.34% reduction in the encoding time for different test sequences at the expense of increasing bitrate up to 0.12% and little decrease in

Table 2. Encoding performance measures in terms of $\Delta PSNR$, ΔR_T and ΔR_B of the proposed algorithm

Sequence Name		Kendo	Balloons	Poznan	Street	Dancer	Average
QP = 24	ΔR_T	43.03%	34.30%	24.41%	63.05%	41.20%	
	ΔR_B	0.29%	0.14%	0.06%	0.42%	0.23%	
	$\Delta PSNR$	-0.08	-0.05	-0.02	-0.22	-0.09	
QP = 28	ΔR_T	43.42%	36.19%	23.51%	60.33%	40.86%	
	ΔR_B	-0.08%	-0.10%	0.35%	-0.16%	0.01%	
	$\Delta PSNR$	-0.03	-0.02	0.00	-0.05	-0.02	
QP = 32	ΔR_T	43.50%	36.31%	18.62%	55.68%	38.53%	
	ΔR_B	0.02%	-0.28%	-0.65%	-0.36%	-0.32%	
	$\Delta PSNR$	-0.02	-0.01	0.00	-0.05	-0.02	
QP = 36	ΔR_T	43.29%	35.33%	18.23%	54.29%	37.78%	
	ΔR_B	0.23%	-0.40%	0.13%	-0.46%	-0.12%	
	$\Delta PSNR$	0.01	0.01	-0.01	-0.07	-0.01	
Average	ΔR_T	43.31%	35.53%	21.19%	58.34%	39.59%	
	ΔR_B	0.12%	-0.16%	-0.02%	-0.14%	-0.05%	
	$\Delta PSNR$	-0.03	-0.02	-0.01	-0.10	-0.04	

Table 3. Quality of the virtual view sequences based on the proposed algorithm and HEVC in terms of *PSNR*

Sequence Name	Kendo(3 → 4)		Balloons(1 → 2)		PoznanStreet(4 → 5)		Dancer(2 → 3)	
	HM	proposed	HM	proposed	HM	proposed	HM	proposed
QP = 24	36.12	36.12	33.89	33.90	34.96	34.96	33.89	33.88
QP = 28	36.09	36.09	33.81	33.78	34.93	34.92	33.84	33.82
QP = 32	36.03	36.01	33.67	33.69	34.87	34.87	33.74	33.73
QP = 36	35.88	35.88	33.52	33.51	34.82	34.83	33.61	33.68
Average	36.03	36.03	33.72	33.72	34.89	34.90	33.77	33.78

PSNR. The reason for the bitrate degradation in Kendo and Dancer test sequences may come from the fact that both sequences are high motion sequences.

From the decoded depth map sequences, we can synthesize the sequences to virtual views by using the DIBR technique [2]. Table 3 shows *PSNR*'s of the virtual view sequences synthesized from the decoded depth map sequences of the conventional HEVC and the proposed algorithm. Table 3 reveals that the proposed method can make good use of the D-NOSE model since the virtual view sequences of the proposed method have the same quality as that of HEVC.

5 Conclusion

Based on the D-NOSE model, the proposed fast coding unit decision algorithm for compressing depth map in HEVC has been presented. The proposed method can determine the minimum coding unit size required for generating the virtual views without any synthesis errors. Using the allowable range calculated by the D-NOSE model, whenever the distorted depth values of the reconstructed coding units are within the range, the coding unit decision process is terminated to speed-up the encoding process.

Experimental results demonstrate that the proposed coding unit decision algorithm significantly outperforms the one in HEVC with respect to computational time complexity, where the proposed method can achieve 21.19% to 58.34% reduction in the encoding time with slight quality and bitrate performance degradation. Furthermore, the proposed algorithm usually generates a virtual view with the same quality as that of HEVC since depth-distortion is considered when determining the size of coding-unit for compressing a 3-D video.

Acknowledgments. K.-L. Chung was supported by the National Science Council of R.O.C. under Contract NSC98-2923-E-011-001-MY3. W.-N. Yang was supported by the National Science Council of R.O.C. under Contracts NSC101-2218-E-011-001 and NSC101-2218-E-011-005. Y.-H. Huang was supported by the National Science Council of R.O.C. under Contract NSC101-2221-E-228-010.

References

1. Fehn, C., Kauff, P., de Beeck, M.O., Ernst, W., IJsselsteijn, M., Pollefeys, M., Van Gool, L., Ofek, E., Sexton, I.: An evolutionary and optimized approach on 3D. In: Proceedings of International Broadcast Conference, pp. 357–365 (2002)
2. Fehn, C.: Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. In: SPIE Stereoscopic Displays Virtual Reality Syst. XI, vol. 5291, pp. 93–104 (2004)
3. Naemura, T., Kaneko, M., Harashima, H.: Compression and representation of 3-D images. IEICE Transactions on Information and Systems E82-D, 558–565 (1999)
4. Wand, R., Wang, Y.: Multiview video sequence analysis, compression, and virtual viewpoint synthesis. IEEE Transactions on Circuit and System for Video Technology 10, 397–410 (2000)
5. Zeng, H., Cai, C., Ma, K.-K.: Fast mode decision for H.264/AVC based on macroblock motion activity. IEEE Transactions on Circuit and System for Video Technology 19, 491–499 (2009)
6. Huang, Y.-H., Ou, T.-S., Chen, H.: Fast decision of block size, prediction mode, and intra block for H.264 intra prediction. IEEE Transactions on Circuit and System for Video Technology 20, 1122–1132 (2010)
7. Wang, H., Kwong, S., Kok, C.-W.: An efficient mode decision algorithm for H.264/AVC encoding optimization. IEEE Transactions on Multimedia 9, 882–888 (2007)
8. Liu, S., Lai, P., Tian, D., Chen, C.-W.: New depth coding techniques with utilization of corresponding video. IEEE Transactions on Broadcasting 57, 551–561 (2011)
9. Cernigliaro, G., Naccari, M., Jaureguizar, F., Cabrera, J., Pereira, E., Garcia, N.: A new fast motion estimation and mode decision algorithm for H.264 depth maps encoding in free viewpoint TV. In: IEEE International Conference on Image Processing (ICIP), pp. 1013–1016 (2011)
10. Leng, J., Sun, L., Ikenaga, T., Sakaida, S.: Content based hierarchical fast coding unit decision algorithm for HEVC. In: International Conference on Multimedia and Signal Processing, vol. 1, pp. 56–59 (2011)
11. Teng, S.-W., Hang, H.-M., Chen, Y.-F.: Fast mode decision algorithm for residual quadtree coding in HEVC. In: Visual Communications and Image Processing (VCIP), pp. 1–4 (2011)
12. Zhao, L., Zhang, L., Ma, S., Zhao, D.: Fast mode decision algorithm for intra prediction in HEVC. In: Visual Communications and Image Processing (VCIP), pp. 1–4 (2011a)
13. Zhao, Y., Zhu, C., Chen, Z., Yu, L.: Depth no-synthesis-error model for view synthesis in 3-D video. IEEE Transactions on Image Processing 20, 2221–2228 (2011b)
14. Tian, D., Lai, P., Lopez, P., Gomila, C.: View synthesis techniques for 3D video. In: Proc. SPIE, vol. 7443 (2009)

Fast Mode Decision Based on Optimal Stopping Theory for Multiview Video Coding

Hanli Wang^{1,2}, Yue Heng^{1,2}, Tiesong Zhao³, and Bo Xiao^{1,2}

¹ Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

² Key Laboratory of Embedded System and Service Computing, Ministry of Education, Tongji University, Shanghai 200092, China

³ Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada

{hanliwang, 1989yueheng}@tongji.edu.cn, ztiesong@uwaterloo.ca, xiaobo_tj@126.com

Abstract. Optimal stopping theory is developed to achieve a good trade-off between decision performance and decision efforts such as the consumed decision time. In this paper, the optimal stopping theory is applied to fast mode decision for multiview video coding in order to reduce the tremendous encoding computational complexity, with the benefit of theoretical decision-making expectation and predictable decision performance. The characteristics of encoding modes in multiview video coding are studied to derive an optimal stopping theory-based model to early terminate mode decision and thus a fast mode decision algorithm is designed. Extensive experimental results demonstrate that the proposed algorithm can save a great amount of encoding time for multiview video coding and meanwhile keep the compression performance more or less intact.

1 Introduction

Multiview (or multi-camera) videos have attracted much attention in a wide range of multimedia applications, such as three-dimensional television, free-viewpoint television, medical imaging, etc. A multiview video consists of video sequences of the same scenario captured by multiple cameras, but from different angles and locations, resulting in the need to store and/or transmit tremendous amounts of data. This subsequently induces a large amount of inter-view statistical dependencies and redundancies in multiview video. Efficient encoding/compression technique is vital for the success of multiview video. As an extension of H.264/Advanced Video Coding (AVC) [1], Multiview Video Coding (MVC) [2] is designed for efficiently exploring not only temporal but also inter-view redundancies, to provide higher coding performance than the independent mono-view coding. In MVC, besides the Motion Estimation (ME) technique to remove temporal redundancy, the Disparity Estimation (DE) is further employed to remove the inter-view redundancy [3], which is achieved by employing more

sophisticated prediction structures [4] with an example as shown in Fig. 1. In each view, either the IBBP or the Hierarchical B-frame Prediction (HBP) structure [5] is supported. Among all views, the first view (*i.e.*, S0 in Fig. 1, also known as the base view) is coded independently; while for the other views, all frames are predicted with temporal ME and/or inter-view DE. Moreover, as the extension of H.264/AVC, variable-block-size ME and DE are employed in MVC, which achieves significant gains in compression efficiency. However, the coding performance improvement is at the cost of dramatically increased coding complexity due to the temporal/inter-view prediction structure and variable-block-size mode decision. Therefore, it is necessary to design optimization approaches to remove the computational obstacles for MVC.

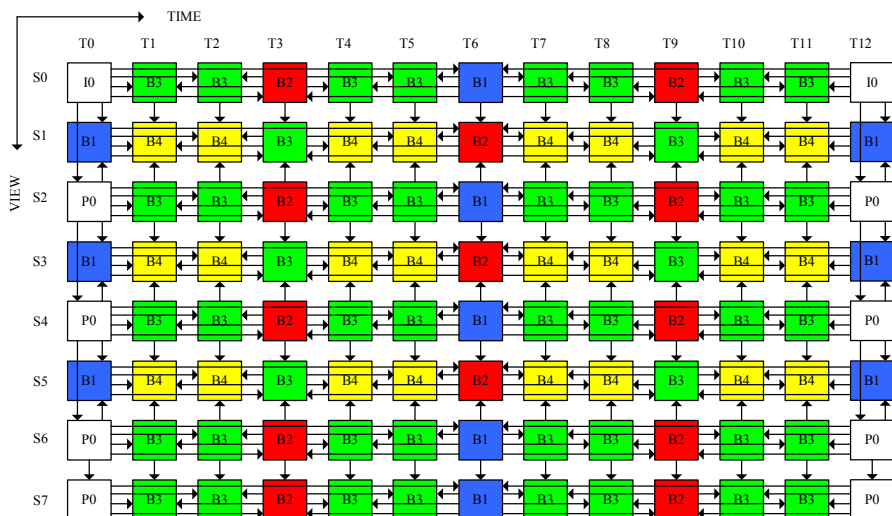


Fig. 1. Multiview video coding with 8 views

To address this issue, a number of researches [6–12] have been investigated on fast mode decision and ME/DE algorithms because most of the coding computations of MVC are consumed by mode decision and the related ME/DE procedures. In almost all these algorithms, the technique of neighboring prediction is applied, by utilizing the neighboring and/or reference information, including but not limited to encoding mode, texture, motion, Rate-Distortion (RD) cost, reference index, and so on. The correlation of encoding modes are studied for fast mode decision such as [6, 11] in which a few encoding modes are checked first and it is determined whether some of the remaining modes can be skipped according to the encoding mode correlation. In particular, fast mode decision based on the SKIP mode is widely applied for early termination of mode decision [7, 8, 10–12]. In these works, the SKIP mode is firstly examined. Then a predefined early termination condition is performed, if the early termination condition holds true, the SKIP mode is considered as the best and thus checking

all the other encoding modes can be skipped. Another major type of fast coding approaches relies on RD cost-based threshold estimation, including [6–12]. These approaches apply early termination to mode decision, multiple reference selection as well as ME/DE processes with termination thresholds derived from correlations of RD costs, motion vectors and disparity vectors.

Despite the above research efforts, fast mode decision for MVC can still be further improved in the following two aspects. Firstly, the mode characteristics and inter-view correlations can be further investigated to reduce the computational complexity. Secondly, the early termination condition for mode decision is usually set empirically, and there is much room to decide the optimal termination condition to skip unnecessary coding modes while keeping the compression efficiency. To this aim, the optimal stopping theory-based model in [13] is further refined and employed in this work for MVC mode decision. Extensive experimental results on benchmark multiview video sequences demonstrate that efficient encoding time reduction and robust decision performance can be achieved by the proposed algorithm. The rest of the paper is organized as follows. In Section 2, based on the investigation of mode characteristics in MVC, an optimal stopping theory-based model is developed. Section 3 provides the overall algorithm with parameter estimation. The experimental results are given in Section 4. Finally, Section 5 concludes this paper.

2 Optimal Stopping Model for MVC Mode Decision

In general, an optimal stopping problem can be defined by a sequence of random variables with known joint distribution and a sequence of real-valued reward functions. The decision-maker examines these variables sequentially to obtain the observed value and decides a time to stop, aiming to maximize the expected reward [14, 15]. By considering the encoding modes as random variables and exploiting joint distribution of modes, the problem of mode decision can be solved with the optimal stopping theory.

2.1 Related Work

In [16], Ferguson *et al.* proposed an optimal stopping problem known as the duration problem, which states that during a decision process, if a variable is with better observed value than any variable before it, then it is called a Relatively Best Object (RBO). The aim of the duration problem is to decide a time to stop with the maximum expected duration until the next RBO. A longer expected duration indicates both a higher probability of no RBO after the current stopping choice and a larger time saving without unnecessary variable examination. In other words, the solution to the duration problem can be considered as a good trade-off between decision accuracy and time reduction, which can be employed to fast mode decision in video coding.

In [13], the solution to the duration problem is investigated for mode decision of scalable video coding and a constrained duration model is developed. In this model, it is assumed there are N candidate encoding modes (which can be regarded as the random variables in the duration problem), denoted as $X_k, k = 1, 2, \dots, N$, and the corresponding probabilities of these modes to be the best are predicted as $\hat{p}_k, k = 1, 2, \dots, N$. With the constraint threshold for decision $\tau \in [N, N + 1)$, all the candidate modes can be ranked in the descending order of probabilities as

$$\hat{p}_i \geq \hat{p}_j, \forall i, j \in [1, N], i < j, \quad (1)$$

where $\hat{p}_k, k = 1, 2, \dots, N$, represent the sorted probabilities. With the rank condition in Eq. (1), the optimal stop for early termination is theoretically derived to be at

$$K_* = \max \{K_\alpha, K_\beta\}, \quad (2)$$

where

$$K_\alpha = \min \left\{ k \geq 1 : \sum_{i=1}^k \hat{p}_i \sum_{j=k}^N \frac{1}{\sum_{r=1}^j \hat{p}_r} > \tau - k \right\}, \quad (3)$$

$$K_\beta = \min \left\{ k \geq 1 : \hat{p}_{k+1} \sum_{j=k+1}^N \frac{1}{\sum_{r=1}^j \hat{p}_r} \leq 1 \right\}. \quad (4)$$

2.2 Mode Characteristics of MVC

In the proposed optimal stopping model as given in Eqs. (2-4), the predicated probabilities $\hat{p}_k, k = 1, 2, \dots, N$, of all candidate modes are utilized to derive the optimal stopping point. In order to make an accurate estimation of these predicated probabilities, the mode characteristics of MVC are investigated. In MVC, the probability of each mode to be the best could be predicted based on the statistical history information and spatial/inter-view correlations. It is highly probable that the co-located MacroBlocks (MBs) in neighboring views have exactly the same best coding mode.

To justify this, for each MB with inter-view prediction being applied to, the intra-view spatially upper, intra-view spatially left, forward inter-view co-located and backward inter-view co-located MBs are separately evaluated in terms of the probability to have the same best mode as that of the target MB. These probabilities are summarized in Table 1, where four benchmark multiview video sequences including *Ballet* and *Breakdancer* in the resolution of 1024×768 as well as *Champagnetower* and *Dog* in the resolution of 1280×960 are tested with 8 views and different Quantization parameters (Q_p); and \parallel denotes the “or” condition. From the table, it can be observed that there exist high probabilities

that these intra-view spatially neighboring and inter-view co-located MBs have the same best encoding mode as the target MB, and the probability gets even higher as Q_p becomes larger. As a consequence, both the intra-view and inter-view correlations of encoding modes can be used for mode probability estimation.

Table 1. Percentage of having the same best encoding mode in MVC

Sequence	Q_p	upper	left	upper left	forward	backward	forward backward
<i>Ballet</i>	20	69.39	69.12	81.05	69.45	68.90	81.30
	36	85.82	85.18	91.23	90.91	90.79	94.77
<i>Breakdancer</i>	20	44.04	44.80	61.10	38.68	39.35	55.19
	36	74.43	73.37	84.17	78.23	76.73	86.66
<i>Champagnetower</i>	20	76.89	76.66	85.78	81.06	81.79	89.66
	36	93.99	93.98	96.50	98.03	97.70	99.04
<i>Dog</i>	20	68.35	65.97	80.04	68.19	70.61	81.31
	36	88.95	86.60	92.70	94.81	94.52	96.73

3 Proposed Overall Algorithm

3.1 Speedup with Fast Mode Decision

To illustrate how much time reduction can be saved by fast mode decision approaches, the best encoding mode for each MB is recorded for several benchmark video sequences and then reloaded under the same coding environment. In such a scenario, the original and reloaded schemes obtain exactly the same coding performance in terms of video quality and compression ratio. However, the encoding computations with the reloaded scheme is dramatically less than the original encoder with two examples shown in Fig. 2, since the mode decision process is not performed by the reloaded scheme.

As observed in Fig. 2, the overall coding time could be significantly saved in all cases, which indicates that there exists a great potential to reduce the MVC encoding computational complexity by skipping unnecessary coding modes or even predicting the best coding mode without the entire mode decision process.

3.2 Model Parameter Estimation

In MVC, there are $N = 6$ types of modes to be decided, as SKIP, 16×16 , 16×8 , 8×16 , P8 $\times 8$ (including 8×8 , 8×4 , 4×8 and 4×4) and INTRA modes (including INTRA 16×16 , INTRA 8×8 and INTRA 4×4). In order to apply the optimal stopping model for early termination of mode decision, the mode probabilities should be estimated as required in Eqs. (2-4). To this aim, an extension of the adaptive statistical method for mode parameter estimation in [13] is employed based on intra/inter-view mode correlations.

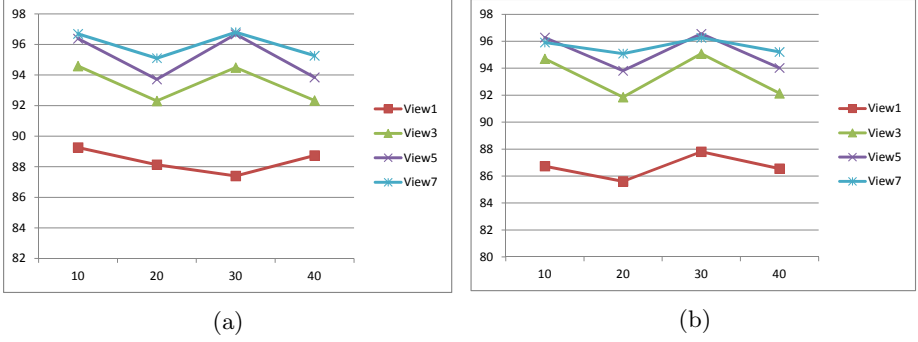


Fig. 2. Two examples of encoding speedup without mode decision. 8 views are encoded with the view coding order 0-2-1-4-3-6-5-7. Y-axis: time reduction (%), X-axis: Q_p . (a) *Ballet* (1024×768). (b) *Dog* (1280×960).

As shown in Table 1, there exists a very high correlation in the best coding mode between an MB and its intra-view neighboring MBs (including intra-view spatially upper MB and intra-view spatially left MB) as well as forward/backward inter-view co-located MBs. Therefore, the information of the best coding modes of the above-mentioned intra/inter-view related MBs is utilized to estimate $\hat{p}_k, k = 1, 2, \dots, N$, with four adaptive prediction matrices being defined as follows.

1. Upper prediction matrix $\mathbf{Tu}(M, k), k = 1, 2, \dots, N, M = 1, 2, \dots, N$, which indicates the percentage of mode k being the best when the best mode of the intra-view upper MB is M ;
2. Left prediction matrix $\mathbf{Tl}(M, k), k = 1, 2, \dots, N, M = 1, 2, \dots, N$, which indicates the percentage of mode k being the best when the best mode of the intra-view left MB is M ;
3. Forward prediction matrix $\mathbf{Tf}(M, k), k = 1, 2, \dots, N, M = 1, 2, \dots, N$, which indicates the percentage of mode k being the best when the best mode of the forward inter-view co-located MB is M ;
4. Backward prediction matrix $\mathbf{Tb}(M, k), k = 1, 2, \dots, N, M = 1, 2, \dots, N$, which indicates the percentage of mode k being the best when the best mode of the backward inter-view co-located MB is M .

With the above prediction matrices, for an MB to be coded, its probability \hat{p}_k as in Eqs. (2-4) could be estimated as

$$\hat{p}_k \approx \frac{\mathbf{Tu}(M_u, k) + \mathbf{Tl}(M_l, k) + \mathbf{Tf}(M_f, k) + \mathbf{Tb}(M_b, k)}{\sum_{r=1}^N (\mathbf{Tu}(M_u, r) + \mathbf{Tl}(M_l, r) + \mathbf{Tf}(M_f, r) + \mathbf{Tb}(M_b, r))}, \quad (5)$$

where M_u, M_l, M_f and M_b represent the best encoding mode of the intra-view upper, intra-view left, forward inter-view and backward inter-view MBs, respectively.

After encoding an MB, if early termination is triggered with Eq. (2), the prediction matrices are updated based on the coding results. If the best mode of the coded MB is obtained as $j \leq K_*$, the posterior probability of mode k to be the best is derived as [13]:

$$\widehat{p}_k = \begin{cases} \sum_{r=1}^{K_*} \widehat{p}_r, & \text{if } k = j, \\ 0, & \text{if } k \leq K_*, k \neq j, \\ \widehat{p}_k, & \text{otherwise.} \end{cases} \quad (6)$$

And the prediction matrices are linearly updated as

$$\mathbf{T}(M, k) = \mathbf{T}(M, k) \cdot (1 - \gamma) + \widehat{p}_k \cdot \gamma, \quad (7)$$

where γ is a regulation parameter with a typical value 0.08 based on exhaustive experiments.

3.3 Overall Algorithm

Based on the above analysis, the proposed overall algorithm is described with the following four steps. In this work, the initial probability for each mode is set as $\frac{1}{6}$, *i.e.*, $\widehat{p}_k = \frac{1}{6}$, $k = 1, 2, \dots, N$ where $N = 6$; and the initial mode checking order is SKIP \rightarrow 16 \times 16 \rightarrow 16 \times 8 \rightarrow 8 \times 16 \rightarrow P8 \times 8 \rightarrow INTRA. In order to achieve a relatively high coding performance, the constraint threshold τ in Eq. (3) is set to be $N + 4/5$.

- Step 1.** Mode probability estimation: compute the mode probabilities \widehat{p}_k , $k = 1, 2, \dots, N$ with intra/inter-view neighboring/co-located MBs as in Eq. (5). Then the optimal stopping point K_* as in Eq. (2) is derived.
- Step 2.** Mode checking order recalculation: reorganize the candidate mode checking list based on the mode probabilities \widehat{p}_k in the descending order, *i.e.*, the larger the probability, the earlier the corresponding mode is to be checked.
- Step 3.** Mode decision: check all the candidate modes according to the checking order (which is derived in Step 2) with optimal stop, and then decide the best coding mode among all examined modes.
- Step 4.** Parameter update: update the posterior probabilities and the related prediction matrices as in Eqs. (6-7).

4 Experimental Results

In order to evaluate the proposed optimal stopping theory-based algorithm, it is implemented on the MVC reference software JMVC with version of 8.3.1. Eight benchmark multiview video sequences are employed for testing, including the 640 \times 480 sequences *Ballroom*, *Exit*, *Race1*, *Vassar*; the 1024 \times 768 sequences *Ballet*, *Breakdancer*, *Doorflowers*; and the 1280 \times 960 sequence *Dog*. The experiment

Table 2. Experimental setting

Encoder	JMVC 8.3.1
Q_p	20, 26, 32, 38
Resolution	640×480, 1024×768, 1280×960
Configuration	Views: 8
	ViewOrder: 0-2-1-4-3-6-5-7
	GOP size: 12
	Frames coded: 8 GOPs (97 frames)
	Number of reference frames: 2
	RDO: enabled
	BiPredIter: 4
	IterSearchRange: 8
ME: fast search with 1/4 pixel	
Search range: ±96	
Entropy coding: CABAC	

Table 3. Experimental results of all 8 views

Sequence	Q_p	TS	Δ PSNR	Δ BR	Sequence	Q_p	TS	Δ PSNR	Δ BR
<i>Ballroom</i>	20	69.51	-0.046	1.98	<i>Exit</i>	20	58.62	-0.035	1.18
	26	65.51	-0.049	2.28		26	68.27	-0.042	2.53
	32	66.81	-0.063	2.51		32	68.62	-0.063	3.45
	38	69.49	-0.090	2.59		38	69.93	-0.100	2.95
	Average	67.83	-0.062	2.33		Average	66.36	-0.060	2.53
<i>Race1</i>	20	63.84	-0.040	1.38	<i>Vassar</i>	20	69.57	-0.026	-0.01
	26	64.21	-0.037	1.44		26	74.13	-0.020	0.69
	32	63.81	-0.046	1.49		32	75.52	-0.037	1.97
	38	63.49	-0.071	1.46		38	73.46	-0.066	2.51
	Average	63.84	-0.049	1.44		Average	73.17	-0.037	1.29
<i>Ballet</i>	20	70.34	-0.031	2.09	<i>Breakdancer</i>	20	56.07	-0.065	3.14
	26	73.12	-0.041	3.07		26	54.31	-0.067	5.10
	32	72.98	-0.073	3.35		32	55.39	-0.088	5.02
	38	71.76	-0.111	3.05		38	55.78	-0.125	3.81
	Average	72.05	-0.064	2.89		Average	55.39	-0.086	4.27
<i>Doorflowers</i>	20	71.21	-0.041	2.14	<i>Dog</i>	20	66.51	-0.019	0.21
	26	73.11	-0.036	3.05		26	62.02	-0.016	0.66
	32	73.12	-0.042	3.02		32	59.56	-0.030	0.74
	38	70.08	-0.056	2.77		38	60.74	-0.059	0.64
	Average	71.88	-0.043	2.75		Average	62.21	-0.031	0.56
Average: TS = 66.59 , Δ PSNR = -0.054 , Δ BR = 2.26									

is executed on a PC of 3.20GHz CPU and 2.00GB memory, with the JMVC configuration parameters given in Table 2.

The proposed algorithm is implemented in all the encoding views with the comprehensive experimental results summarized in Table 3. Three evaluation criteria are used, including TS (%) for time reduction, Δ PSNR (dB) for Peak

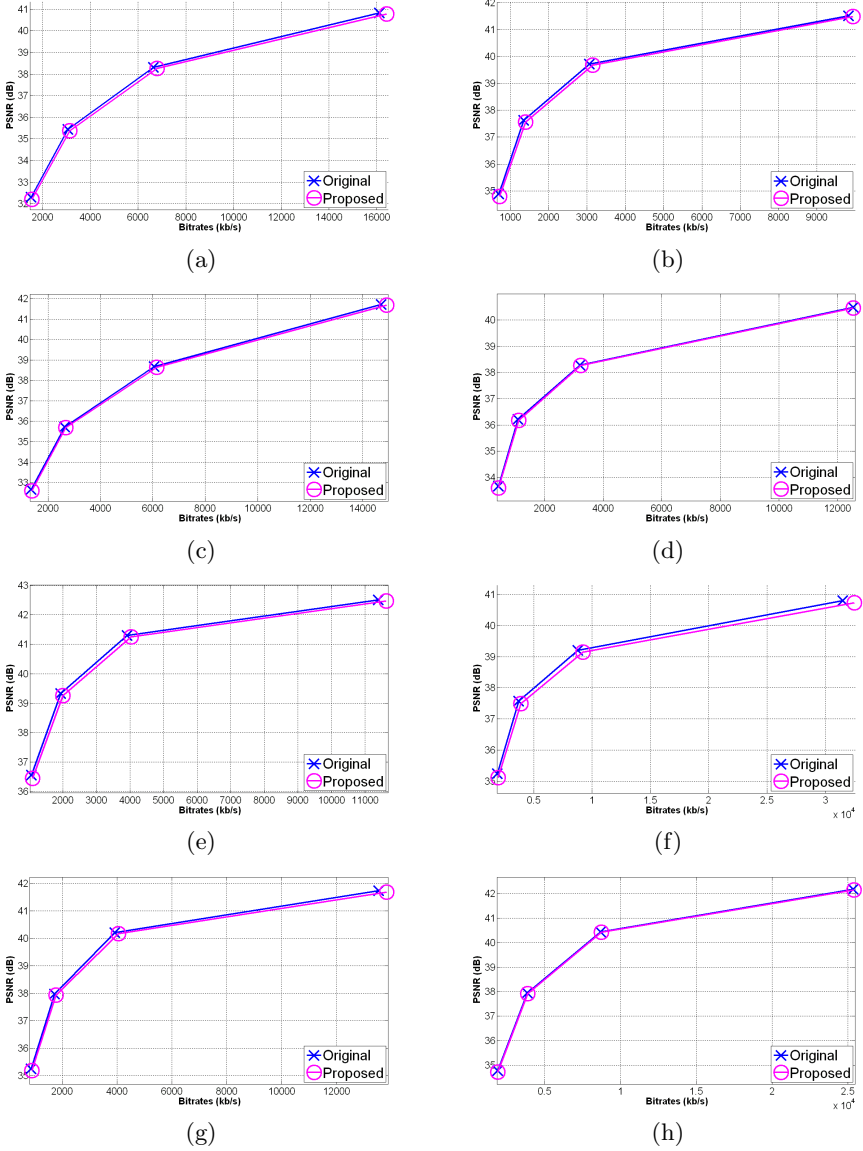


Fig. 3. RD curves. (a) *Ballroom*. (b) *Exit*. (c) *Race1*. (d) *Vassar*. (e) *Ballet*. (f) *Breakdancer*. (g) *Doorflowers*. (h) *Dog*.

Signal-to-Noise Ratio (PSNR) degradation and ΔBR (%) for bitrate increase, as compared to the original encoder, which are defined as:

$$TS = \frac{T_o - T_p}{T_o} \times 100\% \quad (8)$$

$$\Delta PSNR = P_p - P_o \quad (9)$$

$$\Delta BR = \frac{BR_p - BR_o}{BR_o} \times 100\% \quad (10)$$

where T_o , P_o and BR_o stand for the entire encoding time, PSNR and coded bitrate of the original MVC encoder; T_p , P_p and BR_p are the entire encoding time, PSNR and bitrate of the MVC encoder with the proposed algorithm.

For each sequence, the average results of the proposed algorithm are shown in Table 3, and finally the overall average results are presented at the bottom of the table. In average, the proposed algorithm could achieve a significant gain of 66.59% in time reduction with 0.054 dB PSNR degradation and 2.26% bitrate increase, which demonstrates the efficiency and robustness of the proposed algorithm. Moreover, the RD curves are illustrate in Fig. 3, where it can be observed that the RD performance can be kept almost intact as compared to the original encoder.

5 Conclusion

In this paper, the optimal stopping theory is investigated for MVC fast mode decision. The mode correlations in MVC are employed to estimate the probabilities of each candidate mode being the best, and the estimated probabilities are utilized in the proposed optimal stopping model to derive the early termination condition for mode decision. Extensive experimental results demonstrate that the proposed optimal stopping theory-based algorithm is very efficient in reducing MVC encoding computations while maintaining the video quality and compression ratio more or less intact.

Acknowledgments. This work was supported in part by the Shanghai Pujiang Program under Grant 11PJ1409400, the National Natural Science Foundation of China under Grant 61102059, the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the Program for New Century Excellent Talents in University of China under Grant NCET-10-0634, the Fundamental Research Funds for the Central Universities under Grants 0800219158 and 1700219104, the 2010 Innovation Action Plan of Science and Technology Commission of Shanghai Municipality under Grant 10DJ1400300, and the National Basic Research Program (973 Program) of China under Grant 2010CB328101.

References

1. ISO/IEC 14496-10:2005(E) ITU-T Rec. H.264(E): Advanced Video Coding for Generic Audiovisual Services (2005)
2. Vetro A., Pandit P., Kimata H., Smolic A., Wang Y.-K.: Joint Draft 8.0 on Multiview Video Coding. JVT-AB204 (2008)

3. Flierl, M., Mavlanak, A., Girod, B.: Motion and Disparity Compensated Coding for Multiview Video. *IEEE Trans. Circuits Syst. Video Technol.* 17(11), 1474–1484 (2007)
4. Merkle, M., Smolic, A., Muller, K., Wiegand, T.: Efficient Prediction Structures for Multiview Video Coding. *IEEE Trans. Circuits Syst. Video Technol.* 17(11), 1461–1473 (2007)
5. Schwarz, H., Marpe, D., Wiegand, T.: Analysis of Hierarchical B-Pictures and MCTF. In: *IEEE ICME 2006*, pp. 1929–1932 (2006)
6. Ding, L.-F., Tsung, P.-K., Chien, S.-Y., Chen, W.-Y., Chen, L.-G.: Computation-Free Motion Estimation with Inter-View Mode Decision for Multiview Video Coding. In: *3DTV 2007*, pp. 1–4 (2007)
7. Shen, L., Liu, Z., Yan, T., Zhang, Z., An, P.: Early SKIP Mode Decision for MVC Using Inter-View Correlation. *Signal Process.: Image Commun.* 25(2), 88–93 (2010)
8. Zeng, H., Ma, K.-K., Cai, C.: Mode-Correlation-Based Early Termination for Multiview Video Coding. In: *IEEE ICIP 2010*, pp. 3405–3408 (2010)
9. Lin, Y.-H., Wu, J.-L.: A Depth Information Based Fast Mode Decision Algorithm for Color Plus Depth-Map 3D Videos. *IEEE Trans. Broadcast.* 57(2), 542–550 (2011)
10. Shen, L., Liu, Z., An, P., Ma, R., Zhang, Z.: Low-Complexity Mode Decision for MVC. *IEEE Trans. Circuits Syst. Video Technol.* 21(6), 837–843 (2011)
11. Zeng, H., Ma, K.-K., Cai, C.: Fast Mode Decision for Multiview Coding Using Mode Correlation. *IEEE Trans. Circuits Syst. Video Technol.* 21(11), 1659–1666 (2011)
12. Zhang, Y., Kwong, S., Jiang, G., Wang, X., Yu, M.: Statistical Early Termination Model for Fast Mode Decision and Reference Frame Selection in Multiview Video Coding. *IEEE Trans. Broadcast.* 58(1), 10–23 (2012)
13. Zhao, T., Kwong, S., Wang, H., Kuo, C.-C.J.: H.264/SVC Mode Decision Based on Optimal Stopping Theory. *IEEE Trans. Image Process.* 21(5), 2607–2618 (2012)
14. Gilbert, J., Mosteller, F.: Recognizing the Maximum of a Sequence. *J. Amer. Statist. Assoc.* 61(313), 35–73 (1966)
15. Ferguson, T.S.: *Optimal Stopping and Applications*, <http://www.math.ucla.edu/~tom/Stopping/Contents.html>
16. Ferguson, T.S., Hardwick, J.P., Tamaki, M.: Maximizing the Duration of Owning a Relatively Best Object. *Contemp. Math.* 125, 37–57 (1992)

Inferring Depth from a Pair of Images Captured Using Different Aperture Settings

Yujun Li, Oscar C. Au, Lingfeng Xu, Wenxiu Sun, and Wei Hu

Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
{liyujun, eeau, lingfengxu, eeshine, huwei}@ust.hk

Abstract. Given two pictures of the same scene captured using the same camera and the same lens, the one captured with a large aperture will appear partially blurred while the other captured with a small aperture will appear totally sharp. This paper investigates two possible ways of inferring depth of the scene from such an image pair with the constraint that both pictures are focused on the closest point of the scene. Our first method uses a series of Gaussian kernels to blur the image pair, and in the second method, the image pair will be shrunk to a series of smaller dimensions. In both methods, sharp areas in both images will always stay similar to each other, whereas the areas that appear sharp in one image but blurred in the other will not be similar until they are blurred using a large Gaussian kernel or shrunk to small dimensions. This observation enables us to roughly tell which objects in the scene are closer to us and which ones are farther away. At the end of this paper, we will discuss the limitations of our proposed approaches and some of the directions for future work.

Keywords: shape from defocus, depth estimation, 3D reconstruction, aperture, image processing, computer vision.

1 Motivation and Related Work

In recent years, digital cameras not only become more affordable, but also provide functionalities that could potentially be utilized and benefit research in computer vision. For example, since both digital single-lens reflex cameras (DSLR) and mirrorless interchangeable lens cameras (MILC) allow users to change the lens according to the type of effect they want to achieve in their photographs, one can easily capture a clear and sharp landscape shot using a small aperture setting, or create a nice portrait image with a soft background using a lens that offers a large aperture setting. We are interested in the relationship between aperture settings and the effects they have on the images being captured, and we would like to explore and utilize this relationship to infer depth and reconstruct the 3D scene.

The problem of estimating depth from defocused images is called *shape from defocus* in computer vision. Similar to other problems in computer vision such as stereo matching [1, 4], *shape from defocus* is concerned with reconstructing the 3D

shape from 2D images, and various approaches to this problem have been proposed. In [2], the 3D shape of the scene is inferred from the amount of defocus, which is measured at edges where sharp discontinuities of the input images occur. In [3] where the input images are highly textured, a set of interest operators is first computed and then applied to the blurred input images to infer the 3D geometry.

In this paper, we would like to investigate two possible new ways of inferring depth of the scene from a 2D image pair. In our methods, one of the input images is captured using a large aperture while the other is captured using a small one. We also assume that both images are focused on the closest point of the scene. However, since objects in reality are likely to have smooth and textureless surfaces, we do not assume that the input images are highly textured.

2 Introduction

Aperture settings in cameras are closely related to the depth of field of an image. A large aperture will create a shallow depth of field, while a small one will give us a large depth of field. Objects within the depth of field will be sharp, while those outside of it will appear blurred. As a result, given a pair of images I_A and I_B of the same scene captured using the same camera and the same lens, if I_A is captured using a small aperture and I_B a large aperture, then I_A will appear sharp and clear across the whole image while I_B will have a soft background and appear partially blurred.



Fig. 1. Two pictures of the same scene captured using the same camera and the same lens. I_A is captured using an aperture setting of F/22, while I_B is captured using an aperture setting of F/1.8. Both I_A and I_B are focused on the closest point.

Assume that both I_A and I_B are focused on the closest point of the scene. Then it is easy to notice that there is a relationship between the degree of blurredness of an object and its depth. As can be seen from the example in Figure 1, the farther away an object is, the more blurred it will appear in I_B , which is captured using a large aperture. We hope to take advantage of this relationship so that we will be able to infer the depth of each point of the scene. We will investigate two possible methods of doing so in the next two sections. In the first method, a series of Gaussian kernels will be utilized. In the second method, the pair of images will be shrunk to a series of smaller dimensions.

3 Inferring Depth Using Gaussian Kernels

In order to estimate the depth of an object in the scene, we need to measure its degree of blurredness in I_B . Note that if we blur the image pair simultaneously using Gaussian kernels of different sizes, overall, both I_A and I_B will gradually become more and more similar to each other (Figure 2). In fact, sharp areas in both images will stay similar to each other no matter how much they are blurred. However, areas that look sharp in I_A but blurred in I_B will not look similar to each other until they are blurred with a relatively large Gaussian kernel. The larger the Gaussian kernel is, the more similar they will appear. Therefore, based on the size of the Gaussian kernel that is required to make two areas at the same position of I_A and I_B similar to each other, one can estimate the depth of that position.

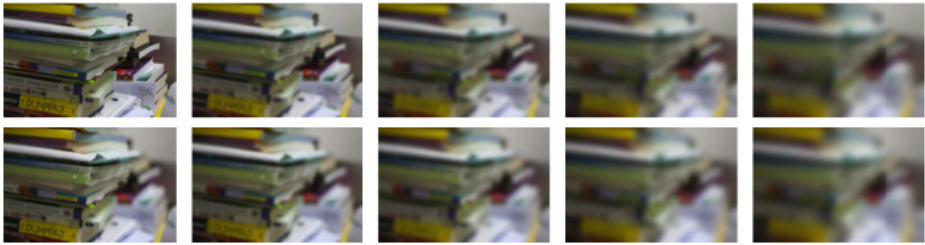


Fig. 2. As the Gaussian kernel we use to blur the image pair becomes larger and larger, I_A (top row) and I_B (bottom row) will gradually become more similar. Sharp areas in both images will stay similar to each other throughout the process. However, areas that are sharp in I_A but blurred in I_B will not appear similar until they are blurred with a relatively large Gaussian kernel. Note that the size of Gaussian kernel increases from left to right.

We start by dividing both images A and B into small blocks (block size: $11\text{px} \times 11\text{px}$) and setting up 40 different Gaussian kernels (h_1, h_2, \dots, h_{40}) ranging from small standard deviation to large standard deviation. For each pair of corresponding blocks W_A and W_B at the same location of I_A and I_B respectively, we compute and store the similarity value between $W_A \otimes h_i$ and $W_B \otimes h_i$ as i goes from 1 to 40. In our experiment, the similarity value between two blocks is computed based on normalized cross correlation (NCC). After that, as i goes from 1 to 40, for each Gaussian kernel h_i , we create a block map M_i that assigns 1 to blocks that have already achieved a similarity value of 0.97 so far and 0 to those that haven't. Therefore, if a block have been assigned 1 when $i = i'$, this block will be assigned 1 for all $i > i'$.

The 40 block maps generated (M_1, M_2, \dots, M_{40}) are shown in Figure 3. As can be seen from the result, sharp areas (such as the corners of books) in I_B achieve the similarity threshold 0.97 earlier in the experiment and therefore show up in the block maps earlier. On the other hand, areas that look sharp in I_A but blurred in I_B (such as the left hand side of the stack of books) achieve the similarity threshold and show up in the block maps later than the sharp areas as i goes from 1 to 40. Overall, the earlier a block shows up in the block map, the closer it is from the camera.

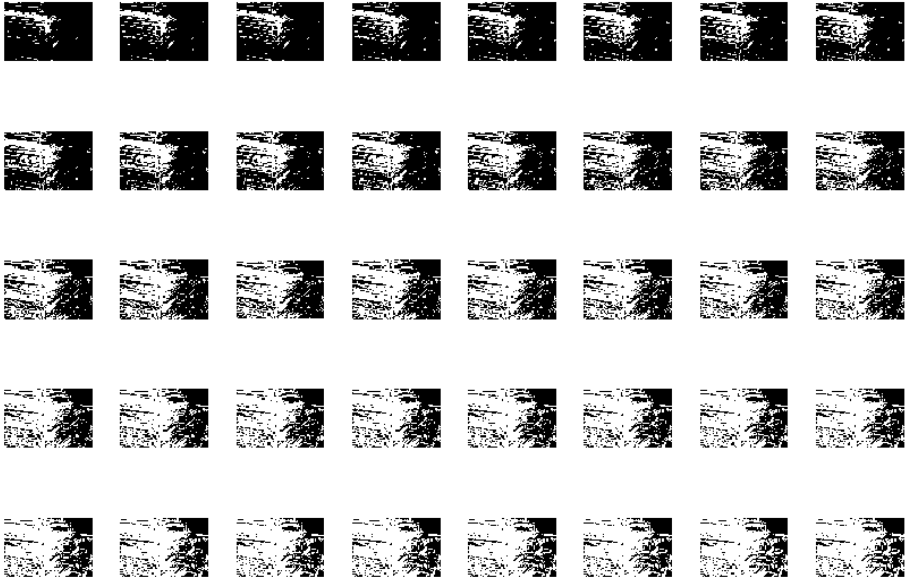


Fig. 3. The 40 block maps (M_1, M_2, \dots, M_{40} from left to right and top to bottom) generated by blurring the image pair using 40 different Gaussian kernels. White pixels in the block maps indicate blocks that have already achieved the similarity threshold of 0.97 so far.

4 Inferring Depth by Shrinking the Image Pair

Besides using a series of Gaussian kernels, we find that shrinking the image pair to a series of smaller dimensions is also useful for estimating depth of the scene. From Figure 4, we can see that as the dimensions of the image pair become smaller and smaller, sharp areas in both I_A and I_B will stay similar to each other no matter how much they are shrunk. However, areas that are sharp in I_A but blurred in I_B will not appear similar until they are shrunk to relatively small dimensions. Based on how much we need to shrink the areas at the same position in I_A and I_B in order to make them similar, we can estimate the depth of that position.

Similar to the previous method where a series of Gaussian kernels are used, I_A and I_B will be divided into small blocks (block size: $11\text{px} \times 11\text{px}$), and 40 different zoom ratios ($r = 1, 1/2, \dots, 1/40$) will be set up for shrinking the image pair. For each zoom ratio r , we first shrink the image pair according to r . Then for each pair of corresponding blocks W_A and W_B at the same location in I_A and I_B respectively, we compute and store the similarity value between the smaller version of W_A and W_B using normalized cross correlation (NCC). Similar to the previous method, as r decreases from 1 to $1/40$, for each Gaussian kernel r , we create a block map M_i that assigns 1 to blocks that have already achieved a similarity value of 0.97 so far and 0 to those that haven't.



Fig. 4. As the image pair becomes smaller and smaller, overall, I_A (top row) and I_B (bottom row) will become more and more similar to each other. While sharp areas in both images will always stay similar to each other, areas that are sharp in I_A but blurred I_B in will not appear similar until they are shrunk to relatively small dimensions.

The 40 block maps generated (M_1, M_2, \dots, M_{40}) are shown in Figure 5. Similar to Figure 3, sharp areas (such as the corners of books) in I_B achieve the similarity threshold 0.97 and show up in the block map earlier than areas that look sharp in I_A but blurred in I_B (such as the left hand side of the book stack) as r decreases from 1 to 40. Blocks that show up earlier in the block maps are generally closer to the camera than blocks that show up later in the block maps.

5 Limitations and Future Work

Even though the proposed methods can roughly estimate the depth of objects in the scene, there are several limitations in our approaches, one of which is that the computation of similarity value is highly dependent on the texture of the image pair. As can be seen from the experimental results in Figure 3 and 5, if an area (such as the white area of the folder at the upper left part of the images) doesn't have much texture, the Gaussian kernels will not be able to significantly alter the appearance of the area in I_A to measure its degree of blurredness. In this case, textureless areas that are blurred in I_B could show up earlier in the block maps.

Similar to the first method where a series of Gaussian kernels is used, lack of texture could also affect the computation of similarity between two corresponding blocks in I_A and I_B in the second method, making blocks that are textureless show up earlier in the block maps.

In order to ensure that blocks in the image pair contain enough texture for measurement of degree of blurredness, one of the directions of our future work is to discard blocks that have little or no texture and consider only the ones that have edges or patterns. However, this could imply that we will need some kind of inpainting or interpolation strategies similar to those used in view synthesis [5] in order to restore

the depth of textureless blocks based on blocks whose depth have been reliably determined using enough texture.

Secondly, the proposed methods do not impose any type of smoothness constraint between blocks. Since depth normally does not change suddenly or abruptly within a single object, we believe that imposing some kind of smoothness constraint can help remove some of the noises and outliers in the block maps.

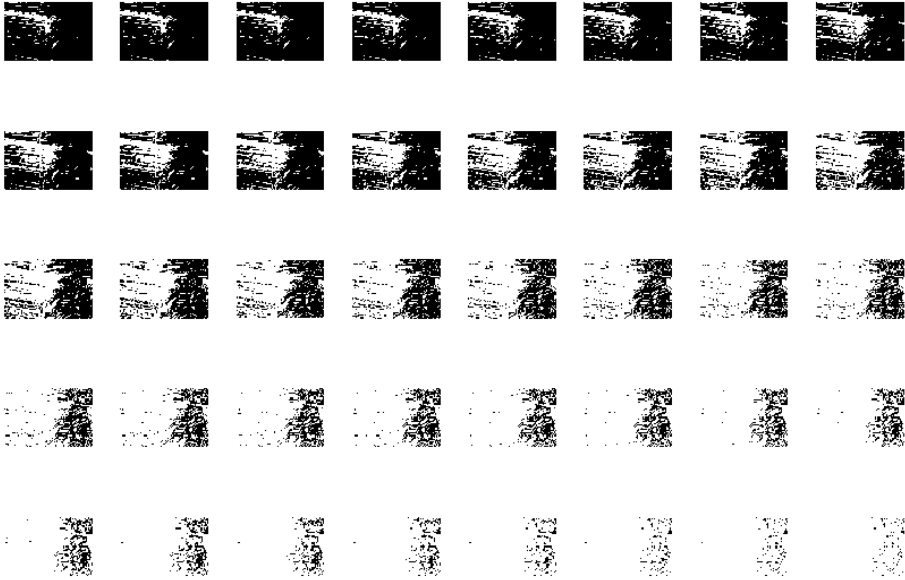


Fig. 5. The 40 block maps (M_1, M_2, \dots, M_{40} from left to right and top to bottom) generated by shrinking the image pair according to 40 different zoom ratios. White pixels in the block maps indicate blocks that have already achieved the similarity threshold of 0.97.

Last but not the least, we would like to establish a more accurate relationship between degree of blurredness and the actual depth. At this point, the blocks maps generated by our methods only allow us to have a rough estimation of depth. To reconstruct the depth map precisely, it is necessary to have a more precise understanding on how to convert depth from degree of blurredness.

6 Conclusion

In this paper, we investigate two possible ways of inferring depth from a pair of images of the same scene captured using the same camera and the same lens. The two images are focused on the closest point in the scene but captured using different aperture settings. Our methods are based on the relationship between degree of blurredness and depth - the farther away a point is, the more blurred it will appear in

the image captured with a large aperture. We measure the degree of blurredness of different areas in the scene by blurring the image pair using a series of Gaussian kernels or shrinking the two images to a series of smaller dimensions and then computing the similarity between two corresponding blocks in the two images. Our methods currently allow us to roughly estimate the depth of scene. In the future, we plan to further improve our methods by inpainting or interpolating the depth of textureless areas, imposing smoothness constraint and more accurately establishing the relationship between depth and degree of blurredness.

References

1. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision* 47(1), 7–42 (2002)
2. Pentland, A.P.: A new sense for depth of field. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (4), 523–531 (1987)
3. Favaro, P., Soatto, S.: A geometric approach to shape from defocus. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(3), 406–417 (2005)
4. Li, Y., Au, O., Xu, L., Sun, W., Chui, S.-H., Kwok, C.-W.: A Convex-Optimization approach to dense stereo matching. In: 2011 IEEE International Conference on Image Processing (IEEE ICIP2011), Brussels, Belgium (September 2011)
5. Oh, K.J., Yea, S., Ho, Y.S.: Hole filling method using depth based in-painting for view synthesis in free viewpoint television and 3-d video. In: 2009 IEEE Picture Coding Symposium (PCS 2009), pp. 1–4 (2009)

Multiple Instance Learning for Automatic Image Annotation

Zhaoqiang Xia, Jinye Peng, Xiaoyi Feng, and Jianping Fan

School of Electronics and Information, Northwestern Polytechnical University,
Xi'an 710072, P.R. China

Abstract. Most traditional approaches for automatic image annotation cannot provide reliable annotations at the object level because it could be very expensive to obtain large amounts of labeled object-level images associated to individual regions. To reduce the cost for manually annotating at the object level, multiple instance learning, which can leverage loosely-labeled training images for object classifier training, has become a popular research topic in the multimedia research community. One bottleneck for supporting multiple instance learning is the computational cost on searching and identifying positive instances in the positive bags. In this paper, a novel two-stage multiple instance learning algorithm is developed for automatic image annotation. The affinity propagation (AP) clustering technique is performed on the instances both in the positive bags and the negative bags to identify the candidates of the positive instances and initialize the maximum searching of Diverse Density likelihood in the first stage. In the second stage, the most positive instances are then selected out in each bag to simplify the computing procedure of Diverse Density likelihood. Our experiments on two well-known image sets have provided very positive results.

Keywords: Automatic Image Annotation, Multiple Instance Learning, Diverse Density, Positive Instance Identification, AP Clustering.

1 Introduction

With the exponential growth of digital images, there is an urgent need to achieve the automatic image annotation for keyword-based image retrieval [7]. For the task of automatic image annotation, machine learning techniques are usually involved to learn the classifiers from large amounts of labeled training images with ground-truth labels corresponded with the content of image regions, which are usually provided by professionals. For the reason that it is labor intensive to hire professionals for labeling large amounts of training images, the size of such professionally-labeled image sets tends to be small. As a result, the classifiers, which are learned from a small set of professionally-labeled training images, may hardly be generalizable.

On the other hand, it is much easier for us to obtain large-scale loosely-labeled training images (object labels are loosely provided at the image level rather than at the region level), which may have multiple advantages: (1) they can represent



Fig. 1. Illustration of multiple instance learning

various visual properties of the object classes more sufficiently; (2) they can be obtained easily by providing the object-level labels loosely at the image level rather than at the region level; (3) both their labels and their visual properties are diverse, thus they can give a real-world point of departure for object detection and scene recognition. Therefore, one potential solution for the critical issue of the shortage of object-based labeled training images is to support multiple instance learning by leveraging large-scale loosely-labeled training images for object classifier training [8,10,13,14].

It is not a trivial task to leverage the loosely-labeled images for object classifier training because they may seriously suffer from the critical issue of *correspondence uncertainty*, e.g., each loosely-labeled image contains multiple image regions (i.e. *instances* in multiple instance learning) and multiple object labels which are given at the image level, thus the correspondences between the regions and the available labels are uncertain. To leverage the loosely-labeled images for object classifier training, it is very attractive to develop new algorithms for: (a) supporting ambiguous image representation which can transform each loosely-labeled image into *bag of instances* and expressing its semantics ambiguity (i.e., multiple labels are available for one single image) explicitly in the instance space; (b) identifying the instance labels automatically when the labels are provided only at the image level; and (c) identifying the positive instances fast for object classifier training. As illustrated in Fig. 1, multiple instance learning tools [1,9,18] can automatically assign multiple labels (which are given at the image level) into the most relevant image regions, where images are also called *bags*. One bottleneck for supporting multiple instance learning is the computational cost on searching and identifying the positive instances in the positive bags. In this paper, a new algorithm of two-stage is developed to speed up the computing of Diverse Density likelihood for multiple instance learning significantly.

The rest of this paper is organized as follows. Section 2 reviews the related work briefly and then the Diverse Density framework our work relied on is presented in Section 3. Section 4 introduces our new algorithm in detail and Section 5 introduces our experimental results for algorithm evaluation. We come to the conclusions in Section 6.

2 Related Work

In the last decades, many multiple instance learning algorithms have been proposed and applied to many fields since the term ‘Multiple Instance learning’ was

created by Dietterich *et al.* [5] in a drug activity prediction domain. The Diverse Density approach has been proposed by Maron [9] to solve multiple instance learning problem and has been improved by Zhang *et al.* [19]. RW-SVM [16] algorithm has used random walk to find out the true positive instances and then used SVM to train the image classifiers to annotate the images. MTS-MLMIL [12] algorithm has utilized graphic clustering to find out the true positive instances and used multi-task SVM to recommend the tags for image annotation. Many other optimization algorithms, such as mi-SVM [1] and MIBoost [15], have been proposed to get the true positive instances through iterative procedures. The other different direction to solve the problem of multiple instance learning is to upgrade the learning process at the instance level (where the labels at the instance level are not available) to the learning process at the bag level (where the bag labels are available), and the classifiers at the bag level are learned for image classification, e.g., recommending the object labels or tags at the image level rather than at the object level.

Some methods compare the bags directly by using the bag-level distance functions, such as Citation-kNN [17]. Chen *et al.* [2] have developed an interesting approach called MILES to enable region-based image annotation when the labels are available only at the image level. Vijayanarasimhan *et al.* have developed a multi-label multiple instance learning approach to achieve more effective learning from the loosely-labeled images [14]. The critical issue for supporting multiple instance learning is how to assign the available labels (which are given at the bag level) into the most relevant instances, e.g., for a given positive bag, some instances are positive and others are negative.

In this paper, we focus on the Diverse Density [9] algorithm for multiple instance learning. The key idea of Diverse Density method is to use the statistical information of all the bags in a probabilistic way, which accumulates the instances to the bag level to utilize the label information which are provided at the bag level. The instances are combined in the Noisy-Or probabilistic model to obtain the likelihood for all the bags. We revisit the Diverse Density algorithm briefly in the next section.

3 Diverse Density

The general framework of Diverse Density uses the likelihood of instances to measure the intersection of the positive bags and negative bags. The *Diverse Density* at one point is defined to measure how many different positive bags have instances near that point and how far the negative instances are away from that point [9]. The task of Diverse Density $DD(x)$ is to find out an appropriate point in the feature space where most true positive instances are around and most negative instances are far away from. That appropriate point in the feature space is called *concept* corresponding with the maximum of the Diverse Density likelihood in the feature space.

In the Diverse Density algorithm, the labeled image sets (which their labels are given at the image or bag level) are defined as D which consist of bags

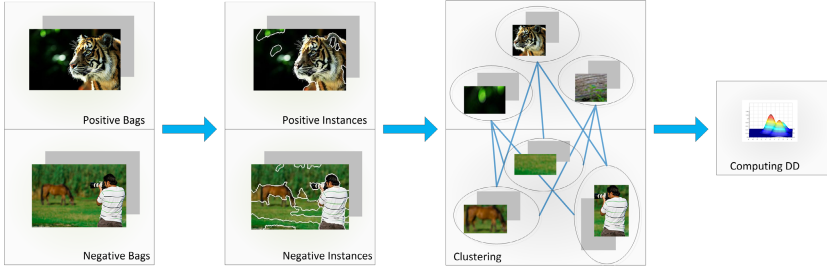


Fig. 2. Illustration of the key components and steps of our proposed algorithm for multiple instance learning

(i.e. images) $B = \{B_1, \dots, B_m\}$ and labels $L = \{l_1, \dots, l_m\}$. Let bag $B_i = \{B_{i1}, \dots, B_{ij}, \dots, B_{in}\}$ where B_{ij} is the j^{th} instance. Let label $l_i = \{l_{i1}, \dots, l_{ij}, \dots, l_{in}\}$ where l_{ij} denotes the label of the j^{th} instance in B_i bag. The positive bags are denoted as B_i^+ and j^{th} instance in B_i^+ as B_{ij}^+ . Likewise, B_{ij}^- represents a negative instance in the negative bag B_i^- . The diverse density at the point t is denoted as $DD(x) = P(x = t | B_1^+, \dots, B_n^+, B_1^-, \dots, B_m^-)$. The maximum point (i.e. concept) can be found out by computing $\operatorname{argmax}_x DD(x)$. Assuming that the candidate point t follows a uniform prior in the feature space, according to the Bayes' rule, this question can be equivalent to

$$\operatorname{argmax}_x DD(x) = \operatorname{argmax}_x \prod_i P(x = t | B_i^+) \prod_i P(x = t | B_i^-) \quad (1)$$

The Diverse Density algorithm uses the noisy-or model to construct the likelihood of positive bags pertained to the candidate point t , $P(x = t | B_i^+) = 1 - \prod_j (1 - P(x = t | B_{ij}^+))$ and the likelihood of negative bags far away from the point t , $P(x = t | B_i^-) = \prod_j (1 - P(x = t | B_{ij}^-))$. Then the distance between the candidate point t and instances is used to model the probability $P(x = t | B_{ij}) = \exp(-\|B_{ij} - x\|^2)$. If the instances in a positive bag are near the t , the probability $P(x = t | B_i^+)$ would be high. Likewise, the probability $P(x = t | B_i^-)$ should be high only if the instances in a negative bag are far away from the candidate point t . To reduce the complexity of Diverse Density algorithm, the negative logarithm of $DD(x)$ can be adopted to search its minimum instead of computing the $DD(x)$ maximum directly.

From (1), one can observe that all the instances need to be integrated into the likelihood $DD(x)$, thus the formula becomes very complicated in computability and could not have the analytic solution while the gradient ascent algorithm can be utilized to search the maximum of $DD(x)$. To avoid trapping a local maximum of $DD(x)$, some initial points need to be adopted to search the global solution at the beginning. So all the positive instances are utilized as its initial points and one of them is likely to be close to the maximum. Although it is beneficial to find out the global solution for the $DD(x)$, it is costly for computing through multiple starting points especially as the number of positive instances are very huge. Based on this observation, our proposed algorithm for multiple instance learning will identify an instance from all these positive instances as the unique

initial point rather than using all the positive instances as the initial point. On the other hand, the $DD(x)$ is actually affected most by the instances which are nearest to the optimal point in each bag. Thus our proposed algorithm for multiple instance learning utilizes only one instance instead of all the instances to maximize the $DD(x)$ in each bag.

4 Our Proposed Method

In this section, we introduce our proposed algorithm for multiple instance learning. Our proposed algorithm consists of 3 key components as shown in Fig. 2: (a) utilizing JSEG[4] segmentation technique to generate instances; (b) utilizing the AP clustering algorithm to find out the best candidate point for Diverse Density likelihood; (c) utilizing the boosting Diverse Density to find out the optimal solution.

4.1 Bag Generation and Instance Clustering

We utilize the JSEG segmentation technique to partition images into a set of regions and all these regions are treated as the image instances, where each image is treated as a bag. For a given category or concept, its relevant images are treated as positive bags and all the irrelevant images are treated as negative bags. The visual features, which are extracted from the instances (i.e. image regions), are used for instance representation. It is worth noting that one object in an image may be partitioned into multiple image regions (multiple instances) or single image region (single instance).

After the instances are generated by using the JSEG segmentation algorithm, these instances would be partitioned into many clusters and only the centers (i.e. exemplars) of these clusters could be selected as the initial points for Diverse Density. The maximum of Diverse Density may locate in the intensive area of positive instances and the non-intensive area of negative instances. In other words, the optimal solution may occur in the vicinity of one of the clusters which gathers the positive instances and does not contain negative instances. Based on this observation, the instances in the positive bags and the negative bags are first partitioned to multiple clusters respectively according to their visual similarity. The cluster in the positive bags, which is farthest away from all the negative clusters in the negative bags, is identified and selected out by analyzing the distances between the positive clusters and negative clusters.

We adopt the AP[6] clustering approach to partition the instances into multiple clusters. The K-means or K-medians clustering approach need to randomly select k initial cluster exemplars at the beginning, thus it is not suitable for the image instance clustering problem because the number of instance clusters is unknown. As an extension of K-medians clustering, AP clustering simultaneously takes all the instances as the potential exemplars, where real-value ‘preferences’ are used for representing the probability of instances as the exemplars. As a result, AP clustering does not need to identify the initial clusters at the beginning and it can automatically detect the exemplars and clusters.

4.2 Candidate Identification and Speeding

As discussed above, clustering is utilized to find out the positive cluster furthest away from negative clusters. The clusters grouped from the positive instances are called *positive clusters* Ω and the clusters gathered from negative instances are called *negative clusters* $\overline{\Omega}$. The candidate point for the Diverse Density maximum should be close to the positive cluster furthest away from negative clusters, and such cluster can be identified by analyzing the distance between the positive clusters Ω and negative clusters $\overline{\Omega}$. In this paper, the Hausdorff distance is used to measure the distance between two clusters or two instance sets.

For two instance sets $C_m \in \Omega$ and $C_n \in \overline{\Omega}$, the Hausdorff distance between C_m and C_n is defined that each element in C_m is within Hausdorff distance d of at least one element in C_n and each point in C_n is within Hausdorff distance d of at least one element in C_m . The Hausdorff distance is defined as:

$$\begin{aligned} H(C_m, C_n) &= \max \{h(C_m, C_n), h(C_n, C_m)\} \\ h(C_m, C_n) &= \max_{x_m \in C_m} \min_{x_n \in C_n} d(x_m, x_n) \end{aligned} \quad (2)$$

where x_m and x_n are the elements in the instance sets C_m and C_n . The distance $d(x_m, x_n)$ between two instances can be any distance measurement corresponding to the feature extraction. The distance we adopt will be introduced particularly in the Sect. 5.1.

Let M and N be the number of positive clusters and negative clusters. We use the score γ to represent the distance between a cluster in the positive clusters Ω and all the negative clusters $\overline{\Omega}$. It is defined as $\gamma(C_m) = \min_n H(C_m, C_n)$. Since the maximum of Diverse Density will locate in the vicinity of C_{m^*} which corresponds with the maximum of $\gamma(C_m)$, we can take any instance in the cluster C_{m^*} as the initial point of Diverse Density. For the fastest convergence, we take the exemplar t_{m^*} of C_{m^*} as the initial point.

Through the selection of initial point, we can search a global optimal solution of the Diversity Density. However, we still need to compute with all the instances in the positive bags according to (1). Through the equation above, we can find that the likelihood of instances is the product of each instance and the maximum of this product is mainly influenced by the nearest instance to the point t . In the view of multiple instance learning, there is always one instance in each bag with most contribution to the bag-level labels. Based on these observations, we take out one instance from each bag to represent this bag and reduce the computing complexity of Diverse Density, which is similar with the Maximization-step in the [19]. We choose an instance from each bag like $\underset{j}{\operatorname{argmax}} P(x = t_{m^*} | B_i(j))$.

After we use the most influenced instance B_{ij^*} to present a bag, we denote the conditional probability between the concept and the bags $P(x = t | B'_i) = P(x = t | B_{ij^*})$ and the j^* is determined by equation above. The $P(x = t | B_{ij}$ in our algorithm are defined as $P(x = t | B_{ij}) = \exp(-d(B_{ij}, x))$. $d(B_{ij}, x)$ can be any measurement between two vectors (i.e. instances). In this paper, we use the distance defined in the (4). For the optimization problem above, we use the numerical solution to find out the optimal solution.

4.3 Boundary Demarcation

Even if the optimal point has been identified, we still need to identify the boundary which would separate the positive bags and negative bags. The cross validation method can be used to find out the best distance threshold, but the search range is still too large. So we can set up the minimum and maximum threshold as the upper and lower boundary for the searching. In multiple instance learning, at least one instance is positive and at most all of them are positive in the positive bags. Based on this, we set the minimum of distance threshold where only one instance in each positive bag occurs and the maximum of distance threshold where all the instances in each positive bag occur. It identifies the boundary as following

$$\begin{aligned} T_{min} &= \max_T d(x_{ij}, t_{m^*}) > T, \quad \exists j \in B_i \\ T_{max} &= \min_T d(x_{ij}, t_{m^*}) < T, \quad \forall j \in B_i \end{aligned} \quad (3)$$

When the search range $[T_{min}, T_{max}]$ is confirmed, we adopt the K-fold (K=10) cross validation method to find the best threshold. The range would be partitioned into L parts and use the centers of L parts to validate the accuracy.

The boundary of some categories may overlap in the feature space and the ranking method is used to solve this. For the concept of i th class C_i , the ranking score of new instance are defined $\gamma(x) = \underset{i}{sort}(\frac{d(x, C_i)}{T_i})$. The distance measurement $d(x, C_i)$ is defined by the (4) and T_i is the best threshold selected out by cross validation. The category with the greatest ranking score is given to the instance as its label.

5 Experimental Evaluation

In this section, we describe the details of our experiments, which include the image sets used, instance feature extraction, the baseline methods and results of our experiments.

5.1 Image Sets and Visual Feature

To evaluate our algorithm precisely, two databases with different size are used to verify the effectiveness of our algorithm. MSRC¹ is collected from search engines, which includes 591 images and 23 object classes. Because the ‘horse’ category in the MSRC only has 3 images, we abandon the ‘horse’ category in our experiments. NUS-WIDE [3] is collected from social website Flickr and contains 269,648 images and tags associated with these images originally. In our experiments, we only use the categories which can correspond to the image regions, i.e. object categories. So we pick out 25 object categories (marked as NUS-WIDE(OBJECT)) to evaluate our algorithm.

¹ <http://research.microsoft.com/en-us/projects/objectclassrecognition/>, we use the version 2.0.

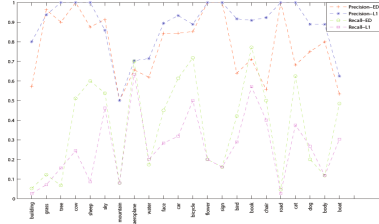


Fig. 3. Precision and Recall of AP clustering on the MSRC datasets based on the ground-truth of instances

To extract effective feature from the instances, we use the colored pattern appearance model (CPAM) to capture both color and texture information of image regions [11]. The CPAM has been applied to many fields and proved succeed to capture the color and texture information. Using CPAM, each instance would be represented by a 64-dimensional feature vector x . The distance between two instances x_m and x_n can be measured by many approaches like L_2 norm of Euclidean distance, however, the L_1 norm has been proved more robust to outliers than L_2 norm. The measurement we adopted is defined as

$$d(x_m, x_n) = \sum_i \frac{|x_m(i) - x_n(i)|}{1 + x_m(i) + x_n(i)} \quad (4)$$

The distance measurement can be used to compute the similarity matrix for the clustering and to determine the threshold for boundary demarcation.

5.2 Experiments

To evaluate the effectiveness of our proposed algorithm, we use the following approaches as our baseline approaches: (a) our algorithm *versus* DD[9] (using multiple positive points as their initial point and search the maximum directly); (b) our algorithm *versus* EM-DD[19] (selecting some positive points and using the most positive instance to compute the maximum) and (c) our algorithm *versus* mi-SVM [1] (another approach to find out the true positive instances using optimization technique directly). For all the approaches mentioned above, we compare the accuracy and running time of algorithms based on the MSRC dataset and NUS-WIDE(OBJECT) dataset. To avoid the over-fitting problem, we randomly generate K training datasets and use the average result of these random datasets to evaluate the effectiveness of our proposed algorithm.

To evaluate the performance of candidate cluster identification, the traditional precision and recall are used to demonstrate the effectiveness of our clustering method. As MSRC database provides the pixel-wise ground-truth images, we utilize these ground-truth images to test our AP clustering performance. As shown in Fig. 3, precision in most categories are high, because most images grouped into one cluster are in the same category. Recall is not very high in many categories because images in the same category are very diverse in the visual

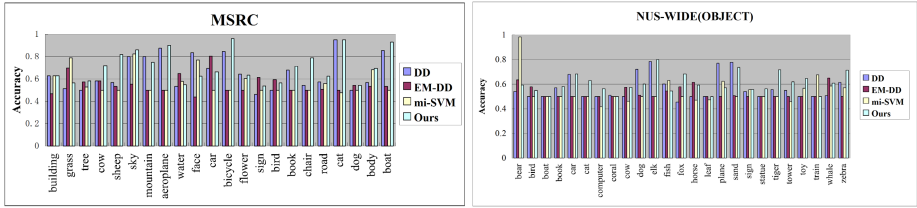


Fig. 4. Average accuracy of image annotation on the MSRC and NUS-WIDE(OBJECT) datasets using four approaches: (a) Diverse Density(DD); (b) EM-DD; (c) mi-SVM; (d) Our algorithm

Table 1. Average accuracy of image annotation on two databases using different methods

	DD	EM-DD	mi-SVM	Ours
MSRC	65.8%	55.2%	56.6%	70.9%
NUS	56.8%	51.7%	50.5%	57.2%

content and not all the positive instances could be contained in one cluster as the cluster can not be infinite. As the positive instances in one category become more and more, the recall would become lower, which can be observed from Fig. 3. As the comparison of our distance measurement, we use the Euclidean distance (ED) as the baseline to observe the effectiveness of different distance measure for the clustering. We can find that the distance defined in (4) can improve the precision with the cost of reducing the recall compared to Euclidean distance from Fig. 3.









To assess the advantages of Diverse Density, we compare the performance between Diverse Density-based approaches and mi-SVM algorithm. As shown in Fig. 4 and Table 1, we can observe from these results that Diverse Density-based algorithms improve the average accuracy for most categories. The improvement of average accuracy is mainly attributed to the fact that Diverse Density algorithms utilize the likelihood of instances to reduce the ambiguity between instances and bag labels. In other words, mi-SVM algorithm utilizes the optimization approach to obtain the relationship between instances and bag labels directly. But it is difficult to achieve that target when the initial labels are not assigned to the right instances at the beginning. In contrast to mi-SVM algorithm, Diverse Density-based algorithms are not necessary to assign the instances with right labels when the algorithms start.

We also compare two existing Diverse Density-based approaches (DD and EM-DD) to evaluate the improvement of our algorithm shown in Fig. 4 and Table 1. The improvement of our algorithm is mainly from 2 components: (a) we use two steps to find out the concept of each category in the feature space. The AP clustering and Hausdorff distance between clusters are utilized to identify the candidate from clusters as the initial point in the first step(coarse step). From

Table 2. Average running time of image annotation on two databases using different methods

	DD	EM-DD(50%)	mi-SVM(100)	Ours
MSRC(s)	780.6	257.1	1.05	1.5
NUS(m)	163.8	61.4	0.2	0.3

Table 3. The results of image-level and object-level annotation

Instance Level					
Image Level	sky,building, grass,road	sky,tree,aeroplane, grass	bird,sky	building,bicycle, road	water,cow, grass
Instance Level					
Image Level	dog	whale	fish,coral	bird,statue	sign,plane,leaf

the initial point, the concept could be found out by using the boosting Diverse Density procedure in the second step(fine step). (b) we design a heuristic method to find out the boundary for the category, which is better than the method only using the leave-one-out approach. The searching range defined in the (3) can avoid the over-fitting problem and speed up the searching speed.

The running time of four algorithms are compared in our experiments shown in the Table 2, which are operated on INTEL Xeon E5420 and Windows 7. The mi-SVM obtains the best performance because many special solution algorithms for SVM have been designed while the Diverse Density-based algorithms need to compute the maximum using the numerical approach. We select out the initial point by clustering and Hausdorff distance to replace the multiple initial point trial in DD and EM-DD, and use one instance of each bag to compute the Diverse Density instead of using all positive instances. So the running time can be saved many order of magnitude. At the same time, mi-SVM can hardly converge stable solution so we need to set the iterative times as 100 for terminating the algorithm manually. Although we do experiments about different number of initial instances for the EM-DD algorithm, we only display the running time using 50% of the positive instances which can make the best performance.

At last, we show some results of our annotation experiments in Table 3, where the images are labeled with image-level and object-level annotations. In the Table 3, the instance-level annotations are shown in the segmented images directly. Some instances (image regions) are not assigned with any tags as if the ranking scores γ of instances are larger than 1.0.

6 Conclusion

In this paper, a novel algorithm is developed to speed up the procedure for *Diverse Density*-based multiple instance learning significantly. In the first, an AP clustering technique is performed on the instances both in the positive bags and the negative bags to identify the candidates and initialize the search of the maximum of the Diverse Density likelihood. And then a boosting Diverse Density algorithm is used to compute the optimal solution. At last many experiments using different methods are performed on two well-known image sets.

Acknowledgment. This work is supported by the doctorate foundation of Northwestern Polytechnical University (No: CX201113) and National Science Foundation of China (under Grant No.61075014 and 60875016).

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: *Advances in Neural Information Processing Systems*, vol. 15, pp. 561–568 (2002)
2. Chen, Y., Bi, J., Wang, J.: Miles: Multiple-instance learning via embedded instance selection. *PAMI* 28(12), 1931–1947 (2006)
3. Chua, T.-S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: *CIVR*, pp. 48:1–48:9. *ACM* (2009)
4. Deng, Y., Manjunath, B., Shin, H.: Color image segmentation. In: *CVPR*, vol. 2. *IEEE* (1999)
5. Dietterich, T., Lathrop, R., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence* 89(1-2), 31–71 (1997)
6. Frey, B., Dueck, D.: Clustering by passing messages between data points. *Science* 315(5814), 972 (2007)
7. Lew, M., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. *TOMCCAP* 2(1), 1–19 (2006)
8. Liu, S., Yan, S., Zhang, T., Xu, C., Liu, J., Lu, H.: Weakly supervised graph propagation towards collective image parsing. *IEEE Transactions on Multimedia* 14(2), 361–373 (2012)
9. Maron, O., Lozano-Pérez, T.: A framework for multiple-instance learning. In: *Advances in Neural Information Processing Systems*, pp. 570–576 (1998)
10. Qi, G., Hua, X., Rui, Y., Mei, T., Tang, J., Zhang, H.: Concurrent multiple instance learning for image categorization. In: *CVPR*, pp. 1–8. *IEEE* (2007)
11. Qiu, G.: Indexing chromatic and achromatic patterns for content-based colour image retrieval. *Pattern Recognition* 35(8), 1675–1686 (2002)
12. Shen, Y., Fan, J.: Leveraging loosely-tagged images and inter-object correlations for tag recommendation. In: *MM*, pp. 5–14. *ACM* (2010)
13. Tang, J., Hong, R., Yan, S., Chua, T., Qi, G., Jain, R.: Image annotation by k nn-sparse graph-based label propagation over noisily tagged web images. *ACM Transactions on Intelligent Systems and Technology* 2(2), 14 (2011)
14. Vijayanarasimhan, S., Grauman, K.: Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization. In: *CVPR*, pp. 1–8. *IEEE* (2008)

15. Viola, P., Platt, J., Zhang, C.: Multiple instance boosting for object detection. In: *Advances in Neural Information Processing Systems*, vol. 18, p. 1417 (2006)
16. Wang, D., Li, J., Zhang, B.: Multiple-Instance Learning Via Random Walk. In: Fürnkranz, J., Scheffer, T., Spiliopoulou, M. (eds.) *ECML 2006. LNCS (LNAI)*, vol. 4212, pp. 473–484. Springer, Heidelberg (2006)
17. Wang, J., Zucker, J.: Solving multiple-instance problem: A lazy learning approach (2000)
18. Yang, J., Yan, R., Hauptmann, A.: Multiple instance learning for labeling faces in broadcasting news video. In: *MM*, pp. 31–40. ACM (2005)
19. Zhang, Q., Goldman, S.: Em-dd: An improved multiple-instance learning technique. In: *Advances in Neural Information Processing Systems*, vol. 14, pp. 1073–1080 (2001)

Combining Topic Model and Relevance Filtering to Localize Relevant Frames in Web Videos

Lei Yi¹, Haojie Li^{1,*}, and Shi-Yong Neo²

¹ School of Software, Dalian University of Technology

² KAI Square Pte Ltd.

yilei@mail.dlut.edu.cn, hjli@dlut.edu.cn, neo@kaisquare.com

Abstract. Numerous web videos associated with rich metadata are available on the Internet today. While such metadata like video tags bring us facilitations and opportunities for video search and multimedia content understanding, some challenges also arise due to the fact that those video tags are usually annotated at the video level while many tags actually only describe parts of the video content. Thus how to localize the relevant parts or frames of web video for given tags is the key to many applications and research tasks. In this paper we propose to combine topic model and relevance filtering to localize relevant frames. Our method is designed in three steps. First we apply relevance filtering to assign relevance scores to video frames and a raw relevant frame set is obtained by selecting the top ranked frames. Then we separate the frames into topics by mining the underlying semantics using Latent Dirichlet Allocation and use the raw relevance set as validation set to select relevant topics. Finally, the topical relevances are used to refine the raw relevant frame set and the final results are obtained. Experiment results on real web videos validate the effectiveness of the proposed approach.

Keywords: Topic Model, web videos, kernel density estimation.

1 Introduction

With the rapid development of web technologies, numerous web videos associated with rich metadata are available on the Internet today. These metadatas provide facilitations to users to retrieve and share video corpus, by way of keyword-based video search. While video metadata, like video name, video tags, and so on greatly benefit the development of web videos, some issues also arise. In the current web video portals, videos are annotated with several tags, but usually these tags are tagged at the video level, i.e., they are the description of whole video content. In fact, many tags are only related to some parts of frames in a video. Such fact affects the video retrieval efficiency. For example, a video about the sports news can be tagged with “basketball”, though only a few shots are actually talking about “basketball”. When we use “basketball” as keyword to search, this video may be returned to users. In such scenario, we will need to

* Corresponding author.

watch the whole video to find our interested part. Thus for better user experiencing and more precise video searching, it's demanding to know which time periods or points of the video are actually related to the keywords, i.e., we need localize the relevant frames in a video to a given concept. By localizing relevant frames, some multimedia research tasks can also be benefited. For example, in concept detection or semantic indexing of video corpus, training samples are needed to train a concept classifier. Due to the difficulty in collecting sufficient manual annotations, recently, researchers began to turn to online web resources like *Flicker* or *YouTube* to get weekly annotated training samples [1][2][3]. However, the performance of trained classifiers are dramatically affected by the noisy tagged contents. If we can localize the time period of relevant shots or frames in videos, frames extracted from these time periods can be served as good training samples.

Previous work like [2][3] represent video frames as vectors in a feature space and use non-parametric kernel densities to estimate a relevance score for each frame. This approach is based on the assumption that the relevant frames are clustered in feature space, thus if a frame has short distance to all the positive samples while long distance to negative samples it would be thought to be likely positive, i.e., relevant to given concept. However, due to the well-known semantic gap between low level features and high level semantics, neighbouring samples in feature space are not necessarily semantically consistent, thus there are many irrelevant frames mixed in the filtered results. In this paper, we enhance the relevance filtering performance from the semantic topic mining perspective. This is based on the observations that videos frames can be separated into topics by mining the underlying latent semantics, and frames in same topic are similar while some topics are more relevant to given concept than other topics. Fig. 1 shows two topics for “basketball”. It can be seen that frames in topic of Fig.1 (a) is more relevant to those in Fig.1 (b). Thus by selecting the relevant topics can guide us to further remove the irrelevant frames and identify the relevant ones.

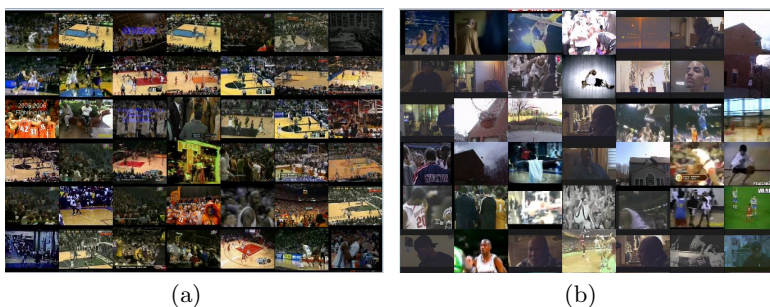


Fig. 1. Two topics assigned to concept “basketball”. While left topic (a) is more likely to relate to the concept. Right topic (b) is less related to the concept.

In this paper, we propose a novel approach to localize the relevant frames in videos. First a probability framework called relevance filtering is applied to modeling content relevance and then top relevant images are selected as validation set to guide the topic selection in topic model. Finally, the selected topics are used to refine the former result and get the final result. Our main contributions are that: (1) we combined the relevance filtering with topic model. (2) We propose a new method to guide topic selection. (3) The result relevant frames are more precise and diverse. We test our approach on a database Youtube22 [2] and shows comparable or moderate improved performance compared to former methods.

2 Related Work

Tag localization on web videos has been studied in recent years. Similar work like web image selection and classification is also being studied. By either identifying the most useful part or removing the noise part, web resources can be best used.

For tag localization, L. Ballan [5] proposed to utilize the social knowledge to suggest and localize tags in web video. With the help of Flickr images, tags in YouTube video can be localized and several tags are suggested to frames in videos. Adrain Ulgs et al. [2] proposed a probabilistic framework for modeling content relevance. This approach is based on a weighted kernel density model [2] and use EM scheme to update the relevance score assigned to each training sample. Based on this approach, they proposed to combine relevance filtering with active learning [3] and used little manual labels to help improve the performance of the model. Manually labeled samples are chosen by several strategies such as random, most relevant, uncertainty and density-weighted repulsion. Another approach also proposed by Adrain Ulgs used relaxed Multiple Instance Learning [4] to solve this problem. Videos are regarded as a bag of key frame-related features. They assumed that training video is weakly labeled. I.e. the associated concept presence variables are observed, but don't know which key frames of a video are "critical". They used MIL-BPNET [6] approach as classifier for bags of samples instead of single samples. By feeding all samples to the network they can obtain a score for each sample. If the score of a sample is within the important fraction, this sample is thought to be relevant.

For web image selection and classification, According to [7][8], most of the existing approaches adopt traditional Euclidean distance or its variants is not enough, also multimodal information need to be considered [9]. So Meng Wang etc.[10] leverages both the visual information of images and the semantic information of tags and propose an diverse relevance ranking scheme to obtain a more relevant and diverse result. Another kind of approach was proposed by the observing that each concept can be divided into several topics, while some topics are more representative, by select such topics we can solve the noise problem. So topic model such as PLSA is used to refine or select most relevant images from web image sets such as [11][12]. One important problem with topic model is how to select the relevant topics. In [12], R. Fergus etc. constructs a validation set by using query represented by six kinds of languages to search in image

search engines and obtain the top few result images. Their assumption is that only the first few images of the first page are likely to be good examples. After the validation set is constructed, it is used to select the most relevant topic among the learned topics by PLSA. Another topic selection approach proposed by Keiji Yanai [11] is similar with [12], while the difference is that in [12] the validation set is obtained by using HTML-txt-based image selection. In [12] only one relevant topic is selected while in [11] several topics can be relevant to the given concept. In this paper, we also produce a validation set using the result of relevance filtering to guide the topic selection.

3 The Proposed Approach

Given a video V and key frames $\{F_i\}$ extracted from it using shot boundary detection algorithms, our goal is to automatically localize the most relevant shots or frames in the video with respect to a certain concept c . To do this, we first estimate a relevance score $P(F_i|V, c)$ for each frame in the video set. Here the positive video set for concept c is constructed by using a set of returned videos $\{V_1, V_2, \dots, V_n\}$ from YouTube by issuing the concept as query keyword; and the negative video samples for concept c are the positive videos of other concepts.

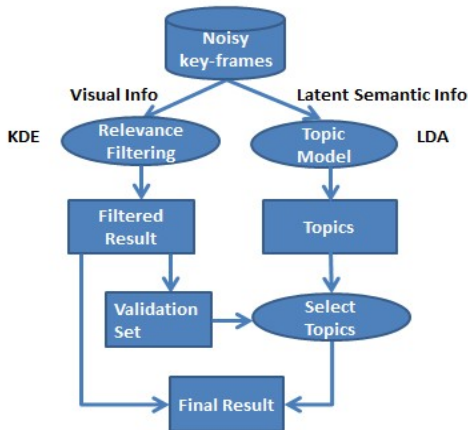


Fig. 2. The overview of the proposed approach for selecting relevant frames from noise key-frames

3.1 Overview

In previous work, images are selected by either feature distance based method [2][3] or topic model based method [11][12]. Feature distance based methods consider image contents while the topic model based methods can model the underlying latent semantic information of images. In this paper we combine the advantages of both to localize the (most) relevant frames in web videos for a

given concept, where kernel density estimation method (KDE) is used for visual relevance filtering and LDA is adopted to model the semantic cues. Figure 2 shows an overview of our proposed approach. First relevance filtering and LDA are applied to the noise key frames to get filtered result and topics respectively. Then use the filtered result as validation set to select relevant topics. Finally, the selected topics can help to refine the filtered result and get the final result.

3.2 Relevance Filtering

Here we adopt the framework proposed in [2] to identify initially relevant frames. In this framework, video frames are coded as vectors in feature space and non-parametric kernel densities are used to model the distribution of relevant and non-relevant frames. The relevance filtering is based on a weighted kernel density model which models the two class-condition densities p^1 (concept presence) and p^0 (concept absence) as follows.

$$\begin{aligned} p_\beta^1 &= \frac{1}{Z'_1} \cdot \sum_{i=1}^n \beta_i \cdot K_h(x; x_i) \\ p_\beta^0 &= \frac{1}{Z'_0} \cdot \sum_{i=1}^n (1 - \beta_i) \cdot K_h(x; x_i) \end{aligned} \quad (1)$$

In the above equation, θ_i is the relevance score for sample x_i , meaning the probability of being relevant. If θ_i is high, sample x_i will have a strong influence on the density p_β^1 but low influence on the density p_β^0 . $Z'_1 = \sum \beta_i$ and $Z'_0 = n - Z'_1$ are normalization constants. For a kernel function K_h , the well-known Epanechnikov kernel with Euclidean distance function and bandwidth h is used:

$$K_h(x; x') = 0.75 \cdot \left(1 - \frac{\|x - x'\|^2}{h^2}\right) \cdot 1_{\|x - x'\| < h} \quad (2)$$

$1_{\|x - x'\| < h}$ is a indicator function whose value is 1 if the condition $\|x - x'\| < h$ is satisfied, otherwise is 0. The kernel function shows the correlation between x and x_i , the shorter the feature distance is, the more x_i will influence x . To calculate the relevance score β , we start with a β^0 and then update the value β^k to β^{k+1} by following iteration.

$$\begin{aligned} \beta_i^{k+1} &:= \frac{P(y_i = 1 | x_i, \tilde{y}_i = 1)}{P(y_i | \tilde{y}_i = 1) \cdot p(x_i | y_i = 1)} \\ &\approx \frac{\sum_{y \in \{-1, 1\}} P(y_i | \tilde{y}_i = 1) \cdot p(x_i | y_i = y)}{\alpha \cdot p_{\beta^k}^1(x_i)} \\ &\approx \frac{\alpha \cdot p_{\beta^k}^1(x_i)}{\alpha \cdot p_{\beta^k}^1 + (1 - \alpha) \cdot p_{\beta^k}^0(x_i)} \end{aligned} \quad (3)$$

The final β will be obtained until converges. Here α is prior for the relevance fraction $\alpha := P(y_i = 1 | \tilde{y}_i = 1)$. This procedure can identify the positive samples in feature space by assigning high score to sample which closing to positive samples and far away from negative samples, while low score to sample closing to the negative samples in feature space.

3.3 Topic Model

Topic model [11][12][13] are traditionally used for text semantic mining where documents are represented as histogram of words. For images, we represent each image in a bag-of-feature manner. First we segment the image into grids and for each grid, the color and texture feature are extracted. Then k-means algorithm is used to generate the cluster centers as codebooks. By assigning the feature of each grid to the closest center, each image is represented as a histogram of cluster centers. We then use LDA [13] to extract the latent semantic topics as its performance and efficiency have been widely validated [14][15]. LDA is a generative probabilistic model assuming each image is represented as random mixtures over latent topics, where each topic is characterized by a distribution over codebooks. Given a corpus consist of M samples, the following procedure is used to assign each sample to K topics:

1. Draw a multinomial θ over K topics: $\theta \sim Dir(\alpha)$
2. For each topic $k = 1 \dots K$
 - Draw a multinomial $\phi_k \sim Dir(\beta)$
3. For each of the N words w :
 - Choose a topic $z_n | \theta \sim Mult(\theta)$
 - Choose a word $w_n | z_n, \beta \sim Mult(\phi_{z_n})$

where α is the parameter of the Dirichlet prior on the per-image topic distributions, β is the parameter of the Dirichlet prior on the per-topic word distribution, θ is the topic distribution for images and N is the total number of visual words in all images. Given the parameters α and β , the joint distribution of topic mixture θ , a set of N topics z , and a set of N words w is given as:

$$p(\theta, z, w | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta) \quad (4)$$

Learning the various distributions is a problem of Bayesian inference and here we use the Gibbs sampling as an alternative inference techniques. After θ_i for image i is estimated, we can assign image i to topic k with the maximum θ_{ik} .

Note that $\sum_{k=1}^K \theta_k = 1$.

3.4 Relevant Topics Selection

We propose using the relevance filtering results to select relevant topics to the given concept. We first apply the relevance filtering procedure using KDE as described in Section 3.2 to the videoframes, and then select the top N resulted samples which are expected to be more relevant to the concept as a validation set. For each of the K topics z_1, z_2, \dots, z_k , a relevance score $P(pos | z_k)$ is calculated to the concept.

Our method to compute the topic relevance score is based on the assumption that topic which is most likely assigned with the top N samples is supposed to

be most relevant. Thus first for the top N ranked samples, we can get the topic distribution θ as described in Section 3.2 and the probability of image assigned to topic z_k is also available represented as θ_{z_k} . Then for each topic z_k , the score $P(pos|z_k)$ is computed as follows.

$$P(pos|z_k) = \frac{1}{N} \sum_{i=1}^N \theta_i Z_k \quad k = 1 \dots K \quad (5)$$

Topics with higher $P(pos|z_k)$ are selected as relevant topics and we can select one or more topics as relevant, depending on the value of $P(pos|z_k)$.

3.5 Combining Relevance Filtering with Topic Model

After applying the relevance filtering procedure on the training samples, the filtered result is obtained by ranking according to the relevance scores. However, in some cases, some less relevant samples can also get higher score since two classes of densities p^1 (concept presence) and p^0 may be both very small (i.e., the feature distances of these samples to positive samples and negative samples are both larger than h), so the resulted relevance score β (see equation 3) may be very large. Thus the feature distance based method may be insufficient in distinguishing the positive and negative sample. In this paper we propose to use topic model to refine or re-rank the filtering results.

In topic section, we have calculated the probability of being positive $P(pos|z)$ for each topic using the top N samples after relevance filtering. Positive topic here means the latent topic generates images relevant to the given concept while Negative topic means that the latent topic generates irrelevant images. Finally, a topic model based relevance score for image I is computed by marginalizing over topics:

$$P(pos|I) = \sum_{z \in K} P(pos|z)P(z|I) \quad (6)$$

where $z \in K$ represents the latent topic. By combining the relevance score computed by relevance filtering β and topic based relevant score $P(pos|I)$, the final relevance score $P(pos|I)$ can be obtained as follows.

$$S(pos|I) = \lambda\beta_I + (1 - \lambda)P(pos|I) \quad (7)$$

Where λ is a weighting factor to balance the importance between relevance filtering and topic model.

4 Experiments

We evaluated our approach on a real-world web video dataset [2][3] which contains 22 concepts and for each concept 100 video clips are downloaded by querying the video sharing website YouTube.

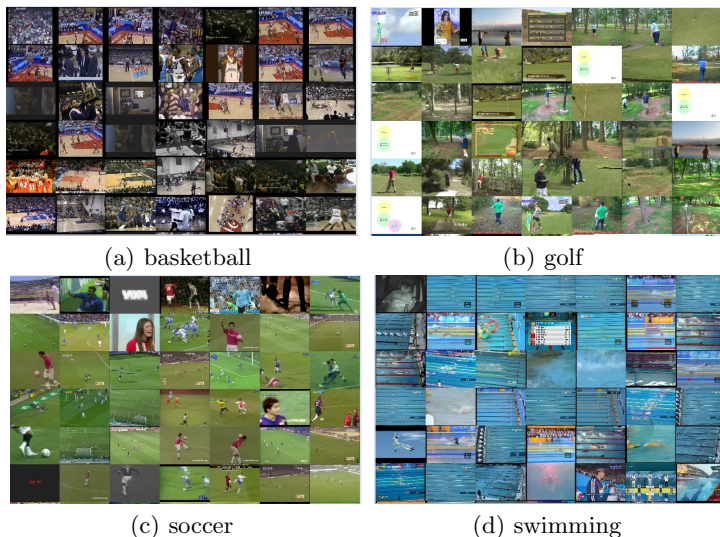


Fig. 3. Top ranked relevant keyframes for four concepts using relevance filtering with KDE: (a) basketball, (b) golf, (c) soccer, and (d) swimming

Similar to [2], we selected 4 concepts including “basketball”, “golf”, “soccer” and “swimming” for evaluation. For each concept, the keyframe samples are extracted from videos using shot boundary detection algorithm, and the noisy positive samples are constructed with samples of the concept, while negative samples are those from other concepts.

For the feature representation of samples, we first divide the sample images into 5×5 grids, and for each patch, the color histogram and edge histogram features are extracted. Then following the bag-of-words approach, we conducted clustering on the two kinds of features respectively, and for each kind of feature a 50 dimensional vocabulary is constructed. Finally, we concatenate the two bag-of-words representations into a 100 dimension feature vector to represent each sample.

In the following three steps were conducted to evaluate and compare the three relevant frames localization strategies. We first test the automatic relevance filtering method using KDE proposed in [2] in Step 1, then test the relevant topic selection strategy based on the filtered results of Step 1 in Step 2. In Step 3, we evaluate the proposed combined filtering approach.

4.1 Step 1: Relevance Filtering Using KDE

We test this method similar to [2][3], only that the feature representation is different. We use grid based color histogram and edge histogram feature in this paper. The bandwidth of h and relevance prior α are selected using cross-validation. The total number of video keyframes is 9602. After the iteration converges, the

relevance score for each sample is obtained and then used to rank the video frames. Fig. 3 shows the top 42 ranked frames for each of the four concepts. We can see most of the resulted keyframes are highly relevant, while some irrelevant frames are also selected.

4.2 Step 2: Relevance Filtering with Topic Selection

We apply LDA to the frames from positive videos as described in Section 3.2. The topic number here is set to 5. After the topic distribution of each frame is obtained, we assign each frame to the topic with maximum probability. Then we select one or more most relevant topics to the given concept by using the relevance filtering result obtained in Step 1 as validation set. Fig. 4 shows the results for topic selection. We can see relevant topics and irrelevant topics are selected correctly.

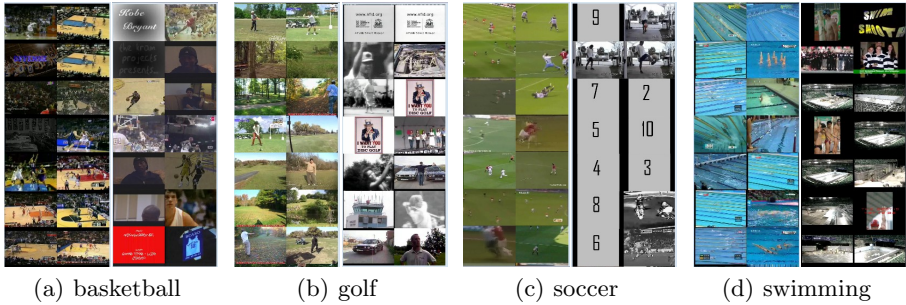


Fig. 4. Topic selection results for concepts: (a) basketball, (b) golf, (c) soccer, and (d) swimming. In each sub-figure, the left column is the most relevant topic while the right one is the most irrelevant topic.

Table 1. Precision on the top 100 and 300 relevant frames returned by automatic relevance filtering [2] and the proposed approach (combining relevance filtering with topic model)

Concepts	P@100 on rel.	P@300 on rel.	P@100 on rel with topic mod.	P@300 on rel with topic mod.
basketball	0.90	0.847	0.98	0.973
golf	0.86	0.926	0.95	0.957
soccer	0.95	0.967	0.99	0.943
swimming	0.94	0.963	0.99	0.99
Mean Precision	0.912	0.926	0.9775	0.967

4.3 Step 3: Combining Topic Model with Relevance Filtering

We refine the relevance filtering results obtained in Step 1 by the topic model. Here, we equally weight the importance of relevance score β and topic model score $p(pos|I)$, i.e., set λ to 0.5. The value of λ can be adjusted according to

different needs. Fig. 5 shows the results of our approach. By comparing Fig. 5 to Fig 3, we can see that some irrelevant frames are removed from top 42 results because their degree of membership in relevant topic is low. Table 1 gives the comparison results for two methods on P@100 and P@300, where P@N is the precision of the top N result frames which is judged manually.

From table 1 we can observe that our approach improves the precision of top results especially for the top 100 images. Meanwhile the P@100 for most of the concepts are lower than their P@300 when using relevance filtering only. Possible reason is that some irrelevant samples may get very high scores when they have little positive density and negative density simultaneously, as explained in Section 3.4. However, when combined with the topic model, such samples are successfully filtered, resulting the relatively higher P@100 and lower P@300.

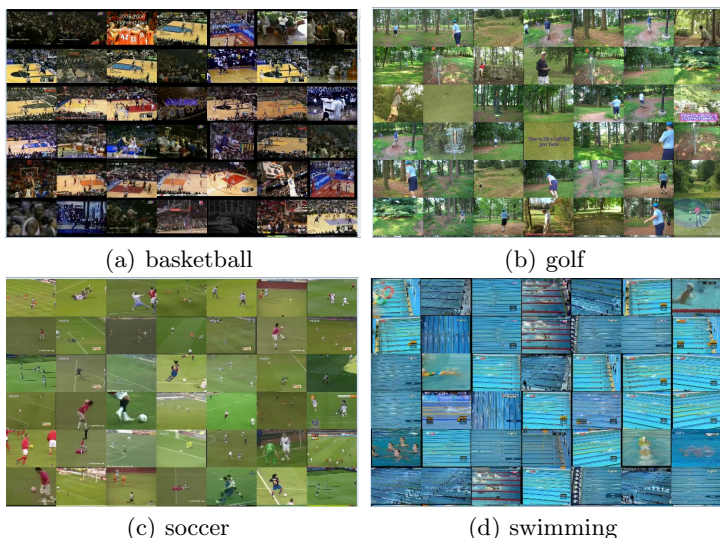


Fig. 5. Top ranked relevant frames for four concepts using the proposed approach: (a) basketball, (b) golf, (c) soccer, and (d) swimming

5 Conclusion and Discussion

In this paper, we aim at automatically localizing relevant frames in web video and a novel approach combining relevance filtering and topic model is proposed. The experiment results show that by integrating the visual information and latent semantic information, we can get better localization performance. However, some issues regarding the proposed approach need further investigation, such as how to determine the topic numbers, the weight balancing relevance filtering and topic selection, and so on. Meanwhile, in this paper, we only consider the case where videos are tagged with single concept. In the future we will study how to mine and leverage the relationships among multiple tags, and how to exploit multi-modality information to localize relevant parts of videos.

Acknowledgements. This work was supported by National Natural Science Funds of China(61033012, 61173104).

References

1. Ulges, A., Schulze, C., Koch, M., Breuel, T.: Learning automatic concept detectors from online video. *Computer Vision and Image Understanding* 114(4), 428–438 (2010)
2. Ulges, A., Schulze, C., Breuel, T.: Identifying Relevant Frames in Weakly Labeled Videos for Training Concept Detectors. In: *Proc. ACM Conference on Image and Video Retrieval* (2008)
3. Borth, D., Ulges, A., Breuel, T.: Relevance Filtering meets Active Learning: Improving Web-based Concept Detectors. In: *Proc. International Conference on Multimedia Information Retrieval* (2010)
4. Ulges, A., Schulze, C., Breuel, T.: Multiple Instance Learning from Weakly Labeled Videos. In: *SAMT Workshop on Cross-Media Information Analysis and Retrieval* (2008)
5. Ballan, L., Bertini, M., Del Bimbo, A., Meoni, M., Serra, G.: Tag suggestion and localization in user-generated videos based on social knowledge. In: *Proc. ACM Multimedia Intl Workshop on Social Media* (2010)
6. Zhang, M.-L., Zhou, Z.-H.: Improve Multi-Instance Neural Networks through Feature Selection. *Neural Process Letters* 19(1), 1–10 (2004)
7. Shen, J., Cheng, Z.: Personalized video similarity measure. *Multimedia Syst.* 17(5), 421–433 (2011)
8. Wang, M., Hua, X.-S., Tang, J., Hong, R.: Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation. *IEEE Transactions on Multimedia* 11(3), 465–476 (2009)
9. Shen, J., Tao, D., Li, X.: Modality Mixture Projections for Semantic Video Event Detection. *IEEE Trans. Circuits Syst. Video Techn.* 18(11), 1587–1596 (2008)
10. Wang, M., Yang, K., Hua, X.-S., Zhang, H.-J.: Towards a Relevant and Diverse Search of Social Images. *IEEE Transactions on Multimedia* 12(8), 829–842 (2010)
11. Yanai, K.: Automatic Web Image Selection with a Probabilistic Latent Topic Model. In: *Proc. of the Seventeenth International World Wide Web Conference, Poster Paper* (2008)
12. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning Object Categories from Google’s Image Search. In: *Proc. of the 10th Inter. Conf. on Computer Vision* (2005)
13. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022 (2003)
14. Wang, C., Blei, D., Fei-Fei, L.: Simultaneous Image Classification and Annotation. In: *Proc. Computer Vision and Pattern Recognition* (2009)
15. Feng, Y., Lapata, M.: Topic Models for Image Annotation and Text Illustration. In: *Proc. Human Language Technologies* (2010)

A Lightweight Fingerprint Recognition Mechanism of User Identification in Real-Name Social Networks

Haibin Cai^{1,2}, Zishan Qin¹, Yunyun Su¹, Junnan Tu¹, and Linhua Jiang^{1,*}

¹ Shanghai Key Laboratory of Trustworthy Computing, East China Normal University, Shanghai, China, 200062

² State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China, 130012

{hbcai, yysu, lhjiang}@sei.ecnu.edu.cn

Abstract. Today, the popularity of social networks poses a great threat to user's information, thus the work of security and privacy protection is becoming increasingly important and urgent. This paper aims to explore the problem of user identification based on biometrics methods in security and privacy issues for social networks sites. In this paper, we propose a lightweight fingerprint recognition mechanism of user identification in real-name social networks. We describe the architecture by using block diagram of our proposed lightweight fingerprint recognition system, and explain how the important steps of proposed mechanism such as minutiae detection, lightweight operation and minutiae matching are implemented. We have performed the experiments to evaluate the user identification reliability of the proposed mechanism. The results of the experiments show that the performance of our lightweight fingerprint recognition system is realistic.

Keywords: Fingerprint recognition, User identification, Security, Social networks.

1 Introduction

Over the past several years, because of the Internet's wide adoption, the popularity of social networks such as the most famous Facebook, MySpace, Twitter, Renren, MSN and Tencent have grown tremendously. Social networks, as virtual communication medium, serve for communication among users, sharing multimedia data, keeping in touch or fun. Therefore, social networks have become critical platform for the information exchange in society's daily life. However, the popularity of social networks poses a great threat to users, and the work of security and privacy protection is becoming increasingly important and urgent [1].

In particular, the real-name system of social networks is encouraged users to create online profiles to represent themselves digitally by providing real name, address, gender, date of birth, school, place of birth, interest and other personal information [2][3]. The sensitive information will be shared with other users and participants can

* Corresponding author.

gain it very easily, however, which can also help attackers in a wide range of networks crimes. Therefore, we must look for proper solution to ensure the user's legitimacy in real-name social networks. In this paper, we focus on the user identification in security and privacy issues for social networks sites. We propose a lightweight fingerprint recognition mechanism of user identification to deal with threats related to illegal users or attackers in real-name social networks. Because of its low-cost and high-accuracy, it is attractive for not only authenticating users but also restricting access to web pages that contain confidential information of real-name social networks application.

The rest of this paper is organized as follows. In section 2, we review user identification related research efforts in the area of social networks. In section 3, we describe the model of user identification system in real-name social networks. In section 4, we propose the lightweight fingerprint recognition mechanism and experiment. In section 5, we describe the experiment results. Finally, in section 6, we conclude the paper and discuss future work.

2 Related Work

User authentication and verification are very important for the security and privacy protection in real-name social networks. The mostly conventional methods of user identification are to adopt user ID and passwords, or personal identification numbers (PINs) [4][18]. Although these conventional methods are easy to implement, there are some disadvantages such that they are easily guessed or forgotten. Therefore, some researchers look for more reliable and friendly solutions to user identification. The biometrics methods is suited for user identification in real-name social networks because that some kinds of biometric characteristics of person such as iris, face, hand-geometry, palm-print etc. are unique, time-invariant and easily observed [5][6][7].

There is a certain amount of related literature to user identification based on biometrics methods over the past few years. Sanches Reillo *et al.* [8] defined and implemented a biometric system based on hand geometry. In identification system they adopt the feature vectors used to describe distances and angles of the hand are the inputs for a comparison process in order to determine the identity of the user whose hand has been photographed. The best results (success rates of approximately 96%) were obtained in this system. Golfarelli *et al.* [9] addressed the problem of performance evaluation in biometric verification systems. In one of two evaluated verification systems they describe the prototype of a hand-based biometric system that takes into account 17 geometrical features of the hand. Jain and Duta [10] presented an authentication method based on the deformable matching of hand shapes. In this method shape distance is automatically computed during the alignment stage, and the systems make the identification decision according to shape distance. The best results (success genuine-accept rates of 96.5%) were obtained in this system. Jain *et al.* [11] presented a prototype of user identification system based on hand geometry. In this prototype of user identification system the identification decision is based on the features of hand geometry including the thickness of the hand and the length and width of the fingers. The best results (a FAR of 0% and FRR of 5%) were obtained in this system. Furthermore, a number of other's works of biometric identification

research such as [12][13][14][15][16][20][21][22] are also very wonderful over the past few years.

3 System Model

The model of our proposed online lightweight fingerprint recognition system consists of a server and networks-connected clients (namely identified users). Firstly, the finger images of the clients are inputted and automatically uploaded to the remote server through the networks. Secondly, after received these information, the server start the recognition process and send the result back to the clients. It includes four main aspects namely finger image acquiring, preprocessing, matching and deciding. Figure 1 shows the block diagram of our proposed lightweight fingerprint recognition system. A low-cost scanner is used as the input device. In the preprocessing module, the standard fingerprint image lightweight procedure is applied. The feature extraction module is used for the extraction of fingerprint features represented by the feature vector. In the subsequent matching module the matching procedure between the corresponding vector and the fingerprint template from a database is performed. In the decision module the correlative rules are used in order to establish identity.

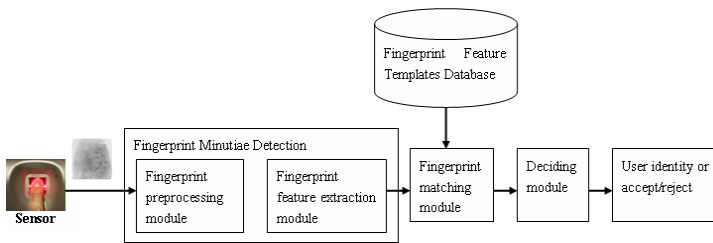


Fig. 1. Block diagram of lightweight fingerprint recognition system

The lightweight fingerprint recognition system has several advantages. Firstly, the user ID regarded as client's account is replaced by fingerprint images. The user's fingerprint cannot be fabricated. Hereby, the identity of users is expected to become more secured. Secondly, in our proposed system, the standard fingerprint image lightweight procedure is helpful to reduce operation time and improve the success rates of user identification.

4 Lightweight Fingerprint Recognition Mechanism and Experiment

In this section it will be explained how the important steps of proposed mechanism and experiment such as minutiae detection, lightweight operation and minutiae matching for a real-name social networks environments are implemented.

4.1 Minutiae Detection

Firstly, an important part of the fingerprint recognition mechanism is the detection of minutiae. This is all contained in one function namely *minutiae_detection*.

A. Estimation of the orientation field.

- First the gradients are calculated for the fingerprint. For this we have used the existing Matlab function *imfilter* with *Sobel kernels*.
- For each *pixel*(i, j), next V_x , V_y , and the orientation (θ) are calculated according to [16].

$$\begin{aligned}
 V_x(i, j) &= \sum_{u=i-W/2}^{i+W/2} \sum_{v=j-W/2}^{j+W/2} 2G_x(u, v)G_y(u, v) \\
 V_y(i, j) &= \sum_{u=i-W/2}^{i+W/2} \sum_{v=j-W/2}^{j+W/2} G_x^2(u, v)G_y^2(u, v) \\
 \theta(i, j) &= (\tan^{-1}(V_x(i, j)/V_y(i, j)))/2
 \end{aligned} \tag{1}$$

where W is the size of the local window; $\theta(i, j)$ is the orientation function; G_x and G_y are the gradient magnitudes in x and y directions; $V_x(i, j)$ and $V_y(i, j)$ are the vector at x and y coordinate axes, respectively.

- Last step is the calculation of the consistency level. These values are calculated for one window size ($=15 \times 15$) for simplicity. It might be important to note that these values are all calculated with a sliding window, not adjacent windows, so every pixel gets a separate value. A disadvantage of this is that it takes longer to calculate.

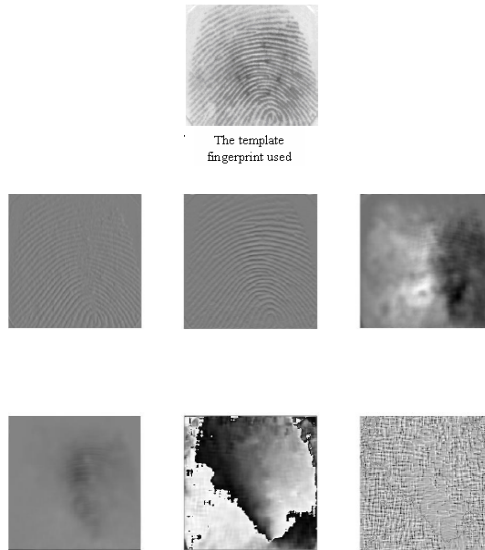


Fig. 2. Results after the step of estimation of the orientation field

The template fingerprint used and the results from these first steps are shown in Figure 2. The first 2 images in the upper left are the gradient images. The next 2 are the values of V_x and V_y and the last 2 images in the bottom right show the orientation

values and the consistency level values. Because of the window operations the size of these images is smaller than that of the original fingerprint. For the pixels in the outer bands of the image one doesn't have a complete neighborhood so values calculated there wouldn't be completely correct.

B. Estimation of the background.

Here we try to remove the background from the fingerprint. The method tries to calculate certainty values of each pixel based on the values of V_x and V_y in formula 2 according to [17]. Figure 3 shows the results. The left image is the values in greyscale. The right image shows the fingerprint where the pixels with a certainty level higher then the threshold value are made black. We found empirically that a value of 40 for the threshold yields good results in most cases. The function responsible for these calculations and of the previous step is *orientation_field*.

$$\begin{aligned}
 H(u, v) &= 2\pi\delta_x\delta_y(\exp(-\frac{1}{2}(\frac{(u-u_0)^2}{\delta_u^2} + \frac{v^2}{\delta_v^2})) + \exp(-\frac{1}{2}(\frac{(v-v_0)^2}{\delta_v^2} + \frac{u^2}{\delta_u^2}))) \\
 f_{enh}(x, y) &= a(x, y)f_{p(x,y)}(x, y) + (1 - \alpha(x, y))f_{q(x,y)}(x, y)
 \end{aligned}
 \tag{2}$$

where u_0 and v_0 are the frequency of a sinusoidal plane wave along the x-axis and y-axis, respectively; δ_x and δ_y are the space constants of Gaussian envelope along x and y axes, respectively; $\delta_u = 1/2\pi\delta_x$ and $\delta_v = 1/2\pi\delta_y$; $f_i(x,y)$ denote the grey level value at *pixel*(x, y) of the filtered image corresponding to the orientation θ_i , $\theta_i = i \times 22.5^\circ$; $p(x, y) = \lfloor \theta(x, y) / 22.5 \rfloor$; $q(x, y) = \lceil \theta(x, y) / 22.5 \rceil \bmod 8$; $\theta(x, y)$ represents the value of local orientation field at *pixel*(x,y); $\alpha(x, y) = \lfloor \theta(x, y) - p(x, y) \rfloor / 22.5$.

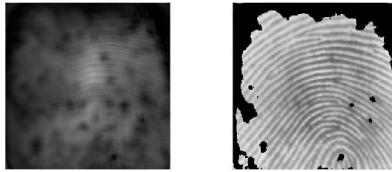


Fig. 3. Results after the step of estimation of the background

C. Detection of the ridges.

For the detection of the ridges, an alternative method according to [18] is implemented which also yields pretty good results but since it performed much slower and its results are slightly worse. It's fairly simple using median filtering, histogram equation and the *bwmorph* function. To see if a pixel is on a ridge, the gray level value of the pixel is compared to the average graylevel of the 8 neighbor pixels. If it is larger, then that pixel is marked as a ridge pixels. The *bwmorph* function is used to fill the holes in the ridges, to reduce the ridges to single pixel width and to remove spurs. The results for the fingerprint in Figure 1 are given in Figure 4 and 5. It can be seen that the ridges in the alternative method are slightly less smooth. This can result in erroneous minutiae being detected. The advantage of both methods is that they don't use the orientation values from the previous calculations like the other method.



Fig. 4. Detected ridges



Fig. 5. Detected ridges with alternative method

D. Detection of the minutiae.

Here also a slightly different approach is taken for some parts. Instead of performing the heuristic for connecting breaks in ridges that are less than 15 pixels large, we remove all endpoint minutiae that lay less than 15 pixels away from each other. The second heuristic, for removing certain branches, is already done in the previous step with the spur removal. To find the minutiae we also use the neighbor counting method on the ridge image as in this paper. For each minutia, the x , y coordinates, the orientation and a sample of ridgepoints from the associated ridge are kept. Minutiae for which the number of associated ridgepoints found is too few are also removed. The function that performs the minutiae detection (*find_minutiae*) also calculates the 1 dimensional representation of the ridge points associated with each minutia.

Figure 6 shows all the candidate minutiae found by the counting method. Squares are candidate endpoint minutiae and stars are candidate bifurcation endpoints. The red lines show the orientation direction for the minutiae that are kept eventually. Figure 7 shows the ridges image with the detected minutiae marked with circles.

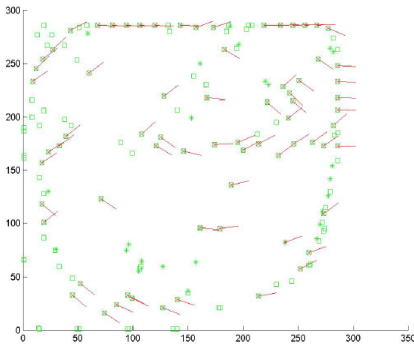


Fig. 6. Detected minutiae



Fig. 7. Detected minutiae superimposed on ridges

4.2 Lightweight Operating

Secondly, another important part of the mechanism is the lightweight operation. After lightweight operating, an original gray image has changed into a binary image, which

makes less data to be saved and processed and enhances contrast between ridge and valley. This is all contained in one function namely *lightweight-operating*.

The first step of lightweight operating is that the segmented image with the crests and valleys will be now binarized. We adopt the method that the black pixels have a value of 1 and white pixels a value of 0. We adopt four rules of the binarization process as follow.

- The image consists only of 0 and 1, where a 1 means a black pixel and a 0 means a white pixel.
- A pixel $I(x, y)$ is considered to be internal, if it's four neighbors $(x+1, y)$, $(x-1, y)$, $(x, y+1)$ and $(x, y-1)$ are $I(blackpixel)$. The limit is defined using 8 neighboring connections.
- A pixel is considered as pixel limit if this isn't an internal pixel and at least one of 8 neighbors is a 0.
- A pixel is considered to be a connection pixel if it is eliminated in a matrix if 3×3 and it's neighbors are disconnected.

After the binarization process, the step of lightweight operating is thinning, in where we find the internal pixels in the image and eliminate the pixel limit. The process as follows is circularly carried out until all internal pixels are picked out.

- The total internal pixels that exist in the image are found.
- All pixels that are a limit pixel are eliminated, but we should take care of condition that this is not a connection pixel.
- The step misapplied again with a small change after thinning the image and finding all internal pixels.
- The last step is again the repetition but in this occasion finding internal pixels with neighbors only.

4.3 Minutiae Matching

The third part of the mechanism is the matching of two minutiae sets in a fingerprint. This is all contained in one function namely *minutiae_matching*.

A. Finding good minutiae pairs

In a first step of the minutiae matching part, all possible minutiae pairs (always one from input fingerprint and one from template fingerprint) are taken and a preliminary score is calculated to see how well the associated ridges correspond to one another. The mechanism uses the one dimensional representation of the ridgepoints for this. The one dimensional representation is obtained by taking the distances for each ridgepoint to the line going through the minutia and in the direction of its orientationfield.

To calculate the score between 2 minutiae/ridges, we have implemented three different formulas as follow:

$$S_1 = \frac{\sum_{i=0}^L d_i D_i}{\sqrt{\sum_{i=0}^L d_i^2 D_i^2}} \tag{3}$$

$$S_2 = \frac{\sum_{i=0}^L \text{abs}(d_i - D_i) / \max(d_i, D_i)}{L-1} \tag{4}$$

$$S_3 = \text{corr}(d, D) \tag{5}$$

where d_j is the distance of the i 'th ridgepoint to the line going through the minutia for input fingerprint; D_i is the distance of the i 'th ridgepoint to the line going through the minutia for template fingerprint; d_i and D_i are the minutiae pairs, total number of pairs is L ; corr is the correlation between the heights.

For the formula 3, we have noticed that the results from this formula don't lay between 0 and 1. Therefore the threshold suggested in the paper couldn't be used. The formula 4 and 5 were added because we noticed that a high score using the first formula didn't necessarily result in a good final matching score. The formula 5 also seemed a good alternative after a brainstorming session. Regretfully also these formulas have the same problem. The approach we have taken is to calculate the scores for all possible pairs and to keep the n best pairs for further processing. We hope that by doing so we at least obtain one good minutiae pair which gives a good transformation in the next step.

B. Aligning the minutiae according to the reference minutiae

This step involves that for each 'good' minutiae pair, the transformation is calculated to go from the reference input minutia to the reference template minutia. This transformation is then applied to all input minutiae. The transformation is calculated according to [16]. This is done for each of the minutiae pairs found in the previous step, resulting in 2 minutiae groups (one for input and one for template) for each pair.

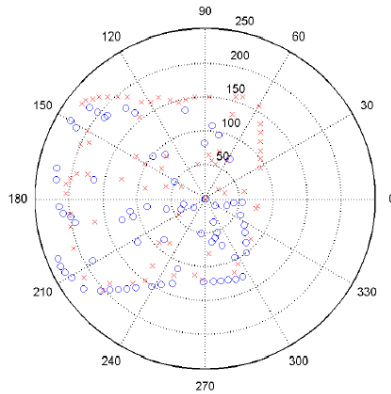


Fig. 8. Aligned minutiae in polar coordinates

C. Transforming to polar coordinates

In this step each of these minutiae groups are transformed to so called strings. This means that first the (x, y) coordinates of the reference minutia of this group are subtracted from the coordinates of all minutiae in the group. This moves the reference minutia to the origin of the XY -plane. Next the (x, y) coordinates of the minutiae are converted to polar coordinates using the *cart2pol* function of Matlab. The minutiae are then sorted according to increasing radial angle. Figure 8 show the results after aligning and transformation to polar coordinates for 2 pairs. As can be that although the matching scores for both minutiae pairs were high, the resulting alignment doesn't necessarily produces good results.

5 Experiment Results

In this Section the correlative results of proposed mechanism and experiment in the real-name social networks environments will be showed.

The online tests were done with database of 100 user's fingerprints of one hundred different people in the real-name social networks, this is, a sample by each person who only is sampled the fingerprint image of index finger.

The tests consisted of the recognition of 150 people, 100 people with stored fingerprint and 50 person with fingerprint not stored. Each person made one tests and we focus on the results demonstrating the user identification reliability.

The programming of the capture of the fingerprint image using a biometric pressure sensitive sensor (BLP-100, BMF Corporation), and the process of transmission by Internet, and the process of fingerprint recognition were made in Matlab. The ROC curve (Receiver Operation Characteristic) of tested mechanism is shown in Figure 9. From the ROC performance result, we can easily see that the black system (with the mechanism in this paper) show great advantage against the red system (without the mechanism introduced in [16]) while maintaining the same FRR (False Rejection Rate) with lower FAR (False Accept Rate).

The time consuming of whole user identification process is evaluated under a normal personal computer with Intel 2.5G Hz CPU and 2G memory, and a networking server of user identification with Quad core (Intel 3.3G Hz CPU) and 6G memory shown as Table 1. From the time consuming of whole user identification process in the real-name social networks, we can easily see that our system (with the mechanism in this paper) show advantage against that system (without the mechanism introduced in [16]), namely, the efficiency is raised by 14.9%.

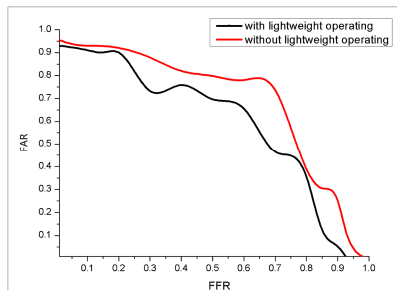


Fig. 9. ROC performance comparison

Table 1. The time consuming of whole user identification process in the real-name social networks

User Identification Process	with the mechanism	without the mechanism
Total Time Consuming	3.37 seconds	3.96 seconds

6 Conclusions and Future Works

In this paper, we have described the lightweight fingerprint recognition mechanism of user identification in real-name social networks. The mentioned mechanism has been designed and implemented in a real-name social networks, and the online tests were be done under the conditions: the programming of the capture of the fingerprint image using a biometric pressure sensitive sensor (BLP-100, BMF Corporation), and the process of transmission by Internet, and the process of fingerprint recognition were made in Matlab. We have performed the experiments to evaluate the user identification reliability of the system. The results of the experiments show that the performance of the lightweight fingerprint recognition mechanism is realistic.

Further work should be undertaken to increase the database size with template files collected, and experimenting under the conditions that a lot of users synchronously log on the real-name social networks sites.

Acknowledgment. This research is partially supported by the Specialized Research Fund for Doctoral Program of Higher Education of China under Grant No.20100076120011 and the Natural Science Foundation of China under Grant No. 91118008 and 61021004. This research is also partly supported by the National High-Tech Research and Development Plan of China under Grant No.2011AA010101, and the State Key Laboratory of Rail Traffic Control and Safety (Beijing Jiaotong University) No.RCS2011K014.

References

1. Gross, R., Acquisti, A.: Information Revelation Privacy in Online Social Networks. In: ACM Workshop on Privacy in the Electronic Society (WPES 2005), pp. 108–121 (2005)
2. Krishnamurthy, B., Wills, C.E.: Characterizing Privacy in Online Social Networks. In: ACM Proceedings of the First Workshop on Online Social Networks (2008)
3. Hay, M., Miklau, G., Jensen, D., Weis, P.: Resisting structural re-identification in anonymized social networks. In: ACM Proceedings of the VLDB (Very Large Data Bases) Endowment (2008)
4. Zheleva, E., Getoor, L.: To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles. In: ACM Proceedings of the 18th International Conference on World Wide Web (2009)
5. Ross, A., Jain, A., Qian, J.Z.: Information fusion in biometrics. In: Proceedings of the 3th AVBA Conference, pp. 354–359 (2001)
6. Frischholz, R.W., Dickmann, U.: BioID: a Multimodal Biometric Identification System. *Computer* 33(2), 64–68 (2000)

7. Shen, J., Cheng, Z.: Personalized video similarity measure. *Multimedia Systems* 17(5), 421–433 (2011)
8. Sanchez Reillo, R., Sanchez Avila, C., Gonzalez Marcos, A.: Biometric identification through hand geometry measurement. *IEEE Transactions Pattern Anal. Mach. Intell.* 22(10), 1168–1171 (2000)
9. Golfarelli, M., Maio, D., Maltoni, D.: On the error-reject trade-off in biometric verification systems. *IEEE Transactions Pattern Anal. Mach. Intell.* 19(7), 786–796 (1997)
10. Jain, A.K., Duta, N.: Deformable matching of hand shapes for verification. In: *Proceedings of IEEE International Conference on Image Processing*, pp. 5–11 (1999)
11. Jain, A.K., Ross, A., Pankanti, S.: A Prototype hand geometry-based verification system. In: *Proceedings of 2th International Conference on Audio- and video-based Personal Authentication (AVBPA)*, pp. 166–171 (1999)
12. Lee, H., Lee, S.H., Kim, T., Bahn, H.: Secure user identification for consumer electronics devices. *IEEE Transactions Consumer Electron* 54(4), 1365–1388 (2008)
13. Zhang, D., Kong, W.K., You, J., Wong, M.: Online Palmprint Identification. *IEEE Transactions Pattern Anal. Mach. Intell.* 25(9), 1041–1050 (2003)
14. Lee, K., Byum, H.: A new face authentications system for memory-constrained devices. *IEEE Transactions Consumer Electron* 49(4), 1214–1221 (2003)
15. Hashimoto, J.: Finger vein authentication technology and its future. In: *2006 Symposium on VLSI Circuits, Digest of Technical Papers*, pp. 5–8 (2006)
16. Wu, J.D., Ye, S.H.: Driver Identification Using Finger-vein Patterns with Radon Transform and Neural Network. *Expert Syst. Appl.* 36(1), 5793–5799 (2009)
17. Jain, A.K., Pankanti, S.: *Automated Fingerprint Identification and Imaging Systems*. In: *Advances in Fingerprint Technology*, 2nd edn. Elsevier Science, New York (2001)
18. Bishnu, A., Bhowmick, P., Dey, J., Bhattacharya, B.B., Kundu, M.K., Murthy, C.A., Acharya, T.: Combinatorial Classification of Pixels for Ridge Extraction in a Grayscale Fingerprint Image. In: *ICVGIP* (2002)
19. Shen, J., Tao, D., Li, X.: Modality Mixture Projections for Semantic Video Event Detection. *IEEE Transactions on Circuits and Systems for Video Technology* 18(11), 1587–1596 (2008)
20. Zhai, L.: The research of double-biometric identification technology based on finger geometry & palm print. In: *Proceeding of 2th International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC 2011)*, pp. 3530–3533 (2011)
21. Conti, V., Militello, C., Sorbello, F., Vitabile, S.: A Frequency-based Approach for Features Fusion in Fingerprint and Iris Multimodal Biometric Identification Systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Application and Reviews* 40(4), 384–395 (2010)
22. Gargouri Ben Ayed, N., Masmoudi, A.D., Masmoudi, D.S.: A new human identification based on fusion fingerprints and faces biometrics using LBP and GWN descriptors. In: *Proceedings of 8th International Multi-Conference on Systems, Signals and Devices (SSD 2011)*, pp. 1–7 (2011)

A Novel Binary Feature from Intensity Difference Quantization between Random Sample of Points

Dongye Zhuang^{1,3}, Dongming Zhang^{1,2}, Jintao Li^{1,2}, and Qi Tian⁴

¹ Advanced Computing Research Laboratory, Institute of Computing Technology,
Chinese Academy of Sciences

² Beijing Key Laboratory of Mobile Computing and Pervasive Device

³ University of Chinese Academy of Sciences

⁴ University of Texas at San Antonio

{zhuangdongye,dmzhang,jtli}@ict.ac.cn, qitian@cs.utsa.edu

Abstract. With the explosive growth of web multimedia data, how to manage and retrieval the web-scale data more efficiently has become a urgent problem, which expects more efficient low-level feature with low computation. This pressing need brings a huge challenge to the conventional feature. It is urgent to make descriptor more compact and faster and meanwhile remain robust to many different kinds of image transformation. To this end, this paper proposed one kind of fast descriptor for local patch. It consists of a string of binary bits which are derived from the intensity difference quantization (IDQ) between pixel pairs which are chosen according to a fixed random sample pattern, so we called it DIDQ (descriptor based on IDQ). Our experiments show that DIDQ is very fast to be computed and also more robust than the other existing binary represented features.

1 Introduction

Along with the explosive growth of web images and videos, visual search application expects more efficient and effective features with low computation complexity and memory consumption. Because of the comparative more discriminant and invariant of image deformation, the local feature has played a control role in visual retrieval application, and even in the field of automatic social multimedia tagging [13], the local feature still are widely used recently. Many kinds of effective local features have been proposed in recent years. Among these features, SIFT (Scale-Invariant Feature Transform) [7] already have been widely accepted and it is highly discriminant and invariant to a variety of image transformations, but with the high expense of computational cost. In order to improve the performance, some derivative works (*e.g.* SURF (Speeded-up Robust Feature) [2] and GLOH (Gradient Location Orientation Histogram) [8]) are proposed, but these approaches still can not finally resolve the conflict between efficiency and effectiveness. These features pay more attention to handle various photometric or geometric transformations, so that they need relatively more computation

resources and could not handle the web scale video database and return the search result in real time. In order to deal with large scale image dataset, Sivic *et al.* [12] proposed Bag-of-Feature (BoF) method which quantize the features to visual words and significantly enhance the efficiency of similar image retrieval, but BoF brings a lot of quantization errors and totally ignores the spatial information. To fix this, Gao *et al.* [4] expand the visual words from a query image for better retrieval recall without the sacrifice of precision and efficiency. To embedding the spatial information, Zhou *et al.* [18] propose a novel scheme spatial coding, to encode the spatial relationships among local features in an image. Zhang *et al.* [16] consider local features in groups to model their spatial contexts. Wang *et al.* [14] and Xie *et al.* [15] proposed semi-local spatial coherent verification(LSC) and pairwise weak geometric consistency constraint (P-WGC) with GPU acceleration to overcome the drawback of BoF. Because of the high bit-rate of low-level features, however, the BoF framework and its derivative works still cannot handle the task of retrieval on the large-scale dataset.

Most recently, for more concise representation and rapid matching, many works which use the binary string to represent a local feature have appeared in the international conferences and journals. Calonder *et al.* [3] proposed the method to combine the FAST [9] keypoint detector and the BRIEF (Binary Independent Elementary Feature) algorithm to detect and describe the local feature with very high speed. But the BRIEF can not handle the change of orientation and scale. Rublee *et al.*[11] promoted the invariant of BRIEF to orientation and proposed the ORB (Oriented BRIEF). Almost simultaneously, BRISK [6] was also proposed to improve the robustness of BRIEF to orientation and scale. FREAK [1] is another effective binary feature, unlike ORB and BRIEF, in which the selection of points are trained by human retina model to improve the performance. In addition to extracting the binary feature from image patch directly, Zhou *et al.* [17] proposed a method to generate the binary feature from the SIFT descriptor. These binary features above are generally compact represented and very high matching speed, but with a considerable loss of discriminant. To summarize, despite having been extensively studied, compact and robust local feature for image still remains a huge challenge in computer vision today.

Generally, effectiveness and efficiency are two competing properties and can not be fulfilled simultaneously. The above binary features could be computed with very high speed due to the compact extraction algorithm and low-bit-rate representation, but the discriminant of these features are much lower than SIFT-like features. In this paper, we propose a novel binary features which not only compute with a very high speed, and also have a adequate discriminant ability. The main properties of our approach and hence our contributions are:

- (i) The descriptor we propose is compact and can be computed very fast.
- (ii) The feature can offer more discriminant than other binary features.
- (iii) DIDQ can obtain high robustness to many image transformation with low computation complexity.

2 Related Works

Both BRIEF and ORB consist of a binary string derived from concatenating the result of simple intensity comparison. The two descriptors can be computed with a very high speed because the feature extracting mainly operate the integer comparison. Furthermore, ORB [11] is rotation invariant because it selects the main direction before extracting the feature of each keypoint. The ORB are bit string description of an image local patch constructed from a set of binary intensity tests. Considering a smoothed image local patch p , a binary test τ is defined by:

$$\tau(p; x, y) := \begin{cases} 0, & p(x) < p(y) \\ 1, & p(x) \geq p(y) \end{cases} \quad (1)$$

Where $p(x)$ is the intensity of the point x . The feature is defined as a vector of n binary tests:

$$f_n(p) := \sum_{1 \leq i \leq n} 2^{i-1} \tau(p; x, y) \quad (2)$$

Many different types of test pairs are considered in [3], and the experiments shows that selecting the points according to Gaussian distribution around the center of patch achieve the best performance. It means that the points which are closer to the keypoint have a higher priority to be selected.

The number of feature bits n can be 128 and 256, and the Hamming distance would be metric to measure the distance between features.

However, in our experiments, we observed that a lot of mismatches appear even when the images do not contain any similar local regions. Figure 1 shows one mismatch example, in which all of matched points are false because the features of the corresponding points are near in Hamming space. What leads to the appearing of so many mismatches? We analysed the feature from the definition of $\tau(p; x, y)$, the difference of a pixel pair is measured by one bit which is determined by the comparison of intensity between two pixels, regardless of the value of difference. So it is reasonable to think too coarse feature quantification should be responsible for these mismatches. Hereinafter, we proposed one novel intensity difference quantization (IDQ) scheme and an improved descriptor based on IDQ (DIDQ) to prove our doubt.

The rest of the paper is organized as follows. In Section 3, we give the details of the proposed DIDQ. The experiment and evaluation are shown in Section 4. Finally, we conclude the paper in Section 5.

3 DIDQ: The Method

Binary features, such as ORB and BRISK usually compose of two stages including *point pairs selection* and *extracting the binary string*. Our method, named DIDQ, involved the same first stage, while in the second stage, we designed two sub-stages, i.e. *Intensity Difference Calculation* and *Difference Quantization*.



Fig. 1. One mismatch example using ORB, almost all the candidate points in the left image are matched with the same one point in the right one

Intensity Difference Calculation. We select pixel pairs according to a fixed pattern derived from Gaussian distribution introduced by Calonder[3]. As shown in the Figure 2, every short black line in the circle represent a random pixel pair, and we select the n couples of points according to a fixed pattern (the Gaussian distribution) in the local patch around the keypoint. And then, we can get a sequence of intensity differences between the points which belong to the same pair. In general we can define the the function $d(n)$ as follows:

$$d(a) = \text{abs}(p(x) - p(y)) \quad (3)$$

Where a is the number of dimension, $p(x)$ is the intensity of the point x , and $\text{abs}(x)$ is the absolute value of x . In Figure 2, there is a sequence of values for each keypoint, and these values can define a unique pattern related to the neighbor area around the keypoint, thus we can use the sequence to represent the keypoint.

Orientation Invariant. We search *the header maximum pattern* to obtain the orientation invariant. After the process of calculating the difference, we have gotten a sequence of value for each keypoint. As Figure 3 shows that, these values can be seen as a circle sequence. Thus, we can always find a unique arrangement for each sequence, which we called *the header maximum pattern*. Here, *the header maximum pattern* is defined as a sequence in which value of the first item is the maximum. Apparently this kind of sequence can define a unique direction, and it is orientation invariant. Our method to get the header maximum pattern is very simple. According to the definition, firstly we find the largest item in the sequence, and then move the header pointer to this item. Figure 3 shows the process of searching the header maximum pattern from a sequence of values.

Intensity Difference Quantization. Generally, the dimensional of feature vector n can be 128 or 256, and each dimension require 8 bits to store, so we are

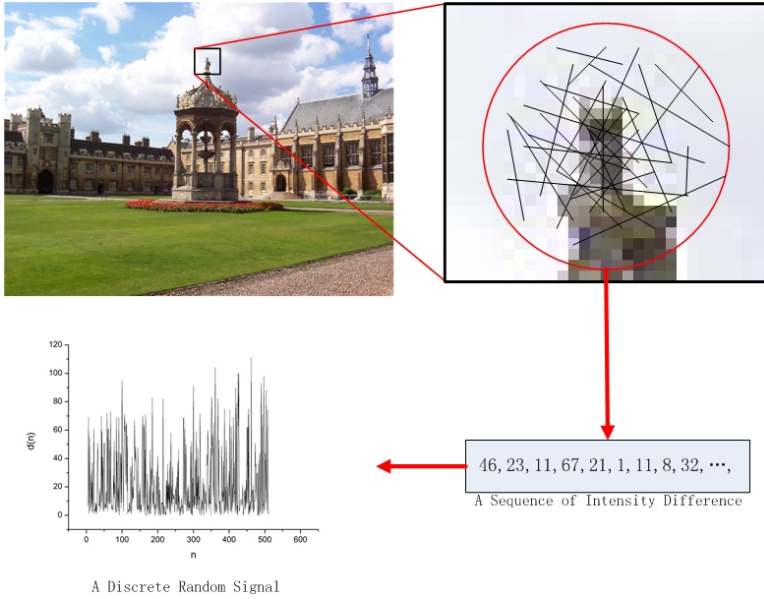


Fig. 2. Intensity difference calculation flowchart. First, We select the pixel pairs according to a fixed pattern generated by 2-D Gaussian distribution[3]. Then we calculate intensity difference for each point pair according to Equation 3.

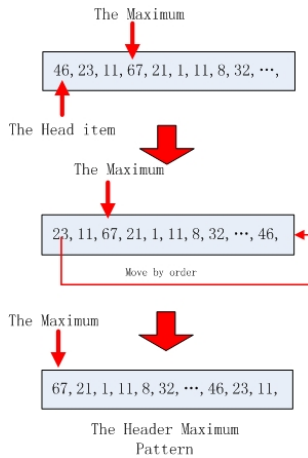


Fig. 3. The process of finding the maximum pattern to obtain the orientation invariant

also confronted with two problems that the bit-rate of this feature is still too high for applications, and furthermore, the distance between two features can not be measured easily. To solve these problems, we should quantize the feature vector to binary bits.

Approximately, the sequence of values $d(x)$ could be seen as a discrete random signal. Here, our quantization target is to classify the signal components into M (in our case, $M = 4$) regions uniformly and this problem can be formulated as follows:

Given the input signal $d(x)$, Classify it into M non overlapping interval $\{R_k\}_{k=1}^M$, by defining $M - 1$ boundary value $\{b_k\}_{k=1}^M$, such that $R_k = [b_{k-1}, b_k)$ for $k = 1, 2, \dots, M$, with the extreme limits defined by $b_0 = 0$ and $b_M = 255$. All the input signal $d(x)$ that fall in a given interval range R_k are associated with the same quantization index k .

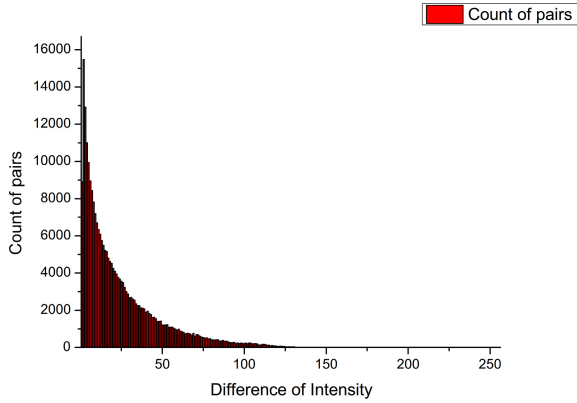


Fig. 4. Distribution of the intensity difference for a image

In our experiments, since $M = 4$, we need to determine 3 boundary value $b_{k=1}^M$. Assuming that the signal d produces random variable X with a associated *probability density function* $f(x)$, the probability P_k that the random variable falls within a particular quantization interval R_k is given by

$$p_k = P[x \in R_k] = \int_{b_{k-1}}^{b_k} f(x)dx \tag{4}$$

Our quantization target is that every p_k should be equal. It means that every signal component falls into a interval region with the same probability. However $f(x)$ is unknown, as a trade-off, we have to use the empirical distribution of the intensity deference as shown in Figure 4 to approach $f(x)$. In Figure 4, we count the number of point pairs which have the same intensity difference, and these point pairs are extracted from the image in Flickr dataset, and we detect 500 keypoints for every image. Apparently, we can approximate it with Gaussian distribution. Herein, non-uniform quantization is a proper choice, which can obtain a uniform distribution after quantization. According to the empirical distribution, we define 4 quantization regions as follows:

$$\begin{cases} R_1 = [0, 10] \\ R_2 = [11, 30] \\ R_3 = [31, 60] \\ R_4 = [61, 255] \end{cases} \quad (5)$$

Next, we can define the select function $S_i(x)$ as follows:

$$S_k(x) = \begin{cases} 1, & x \in R_k \\ 0, & x \notin R_k \end{cases} \quad (6)$$

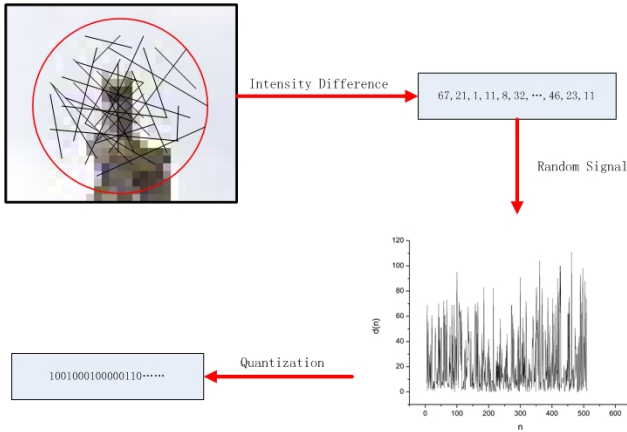


Fig. 5. Quantization for random signal

At last, we can define the quantization function $Q(x)$ as follows:

$$Q(x) = \sum_{k=1}^M y_k S_k(x) \quad (7)$$

Where $M = 4$, $y_1 = '00'$, $y_2 = '01'$, $y_3 = '10'$, $y_4 = '11'$, x is the input signal (difference of intensity), y_k is the bit combination. Thus the quantization transform the discrete signal to a string of bits, and the bit-string is compact, discriminant and orientation invariant. Figure 5 present the quantization process. Therefore, the feature has been defined as a connection of n binary quantization index from Equation 7.

4 Experiment

To verify the effectiveness and efficiency of our method, the dataset [8] are used in our experiments, and it contains 48 images with five different changes in

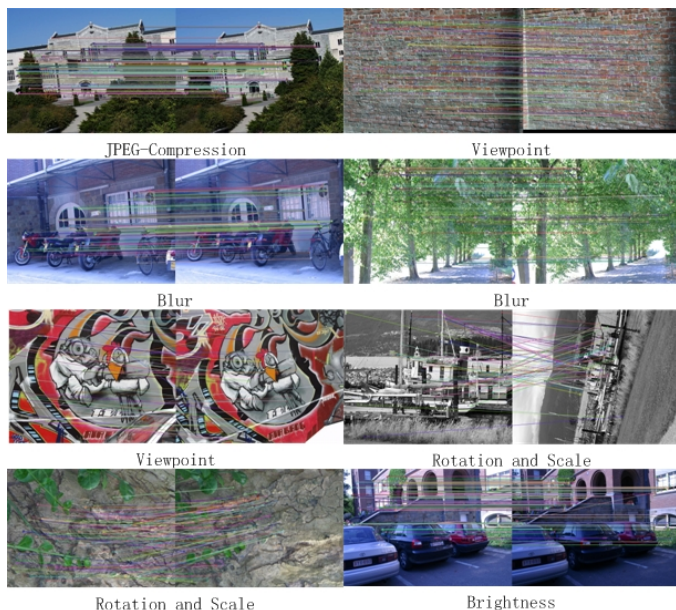


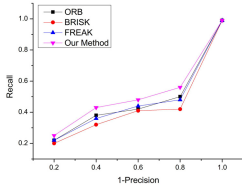
Fig. 6. Image matching result of our method

imaging conditions (viewpoint changes, scale changes, image blur, JPEG compression, and illumination). And it is widely accept that this dataset is appropriate for evaluating the invariant of image transformation. The performance of the proposed method is compared with several binary-represented descriptors (ORB[11], BRISK[6], FREAK[1]). And the result are presented with *recall* versus $1 - \textit{precision}$:

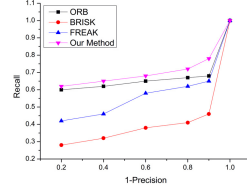
$$\textit{recall} = \frac{\textit{correct matches}}{\textit{correspondences}}, 1 - \textit{precision} = \frac{\textit{false matches}}{\textit{all matches}} \quad (8)$$

Generally, the performance are highly related to the combination of keypoint detector and descriptor. Thus it is necessary to obtain fairness that we present our experiments using the combination of FAST [10] and Harris [5] to detect keypoints, which also used in ORB. Figure 6 shows the demo of image matching by our method with variant transformation. In the experiments, we observed that increasing the number of random sample in one local patch had no impact on the evaluation for the precision, because we found in the experiments that more pairs chosen and more noise generated. Figure 7 presents the quantitative results on the dataset, and it shows our method obtain the best performance.

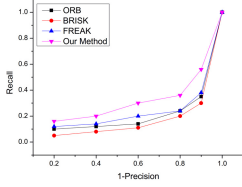
In previous binary features, if the dimensional of feature is n , the number of point pairs is also n . In our method, to obtain the $n - D$ feature, we just sample $n/2$ point pairs, thus our method could significantly reduce the computation complexity in feature extracting process. Accordingly, table 1 presents the computation time of these descriptors, and it shows that our method is faster than



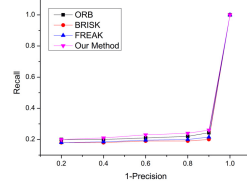
(a) Graf(1-4)



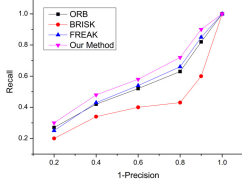
(b) Leuven(1-4)



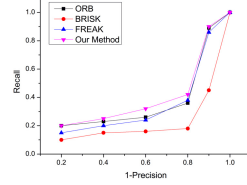
(c) Trees(1-4)



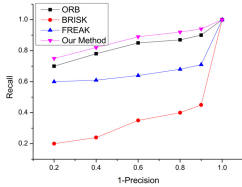
(d) Bark(1-4)



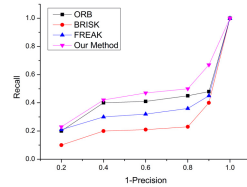
(e) Wall(1-4)



(f) Boat(1-4)



(g) Ubc(1-4)



(h) Bike(1-4)

Fig. 7. Performance evaluation on the dataset introduced by Mikolajczyk and Schmid [8]

Table 1. Computation time on 800×600 resolution images where 1500 keypoints are detected per image. The computation time corresponding to the description and matching of all keypoints.

Time per keypoint	ORB	BRISK	FREAK	Our Method
Description in [ms]	0.011	0.028	0.016	0.009
Matching in[ns]	25	32	25	23

the other binary feature. In the experiment, all algorithms are carried out on PC platform with Intel(R) Core(TM)2 Quad CPU 2.83GHz and 4GB memory.

5 Conclusion

We presented a novel approach to describe the keypoint using a binary string, in which the binary bits derived from quantization of the difference of intensity of pixel pairs. And our experiments showed that DIDQ outperforms the state-of-the-art keypoint descriptors with higher computation speed and lower memory consumption. Therefore, we can conclude that the descriptor DIDQ is very adaptable to web-scale visual search system. In experiments, we found that different point pattern selection in our method may affect the performance. As a future work, we need to investigate how to obtain the optimal selection of pixel pairs to improve the performance.

Acknowledgements. This work is supported by National Nature Science Foundation of China (61273247, 61271428), National Key Technology Research and Development Program of China (2012BAH39B02), and Co-building Program of Beijing Municipal Education, and is supported in part to Dr. Qi Tian by ARO grant W911NF-12-1-0057, NSF IIS 1052851, Faculty Research Awards by Google, NEC Laboratories of America and FXPAL, respectively.

References

1. Alahi, A., Ortiz, R., Vandergheynst, P.: Freak: Fast retina keypoint. In: IEEE Conference on Computer Vision and Pattern Recognition, Rhode Island, Providence, USA, June 16-21 (2012)
2. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded Up Robust Features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
3. Calonder, M., Lepetit, V., Ozuysal, M., Trzcinski, T., Strecha, C., Fua, P.: Brief: Computing a local binary descriptor very fast. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP(99), 1 (2011)
4. Gao, K., Zhang, Y., Luo, P., Zhang, W., Xia, J., Lin, S.: Visual stem mapping and geometric tense coding for augmented visual vocabulary. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3234–3241 (June 2012)
5. Harris, C., Stephens, M.: A combined corner and edge detector (1988)
6. Leutenegger, S., Chli, M., Siegwart, R.Y.: Brisk: Binary robust invariant scalable keypoints. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2548–2555 (November 2011)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 91–110 (2004), doi:10.1023/B:VISI.0000029664.99615.94
8. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(10), 1615–1630 (2005)

9. Rosten, E., Drummond, T.: Fusing points and lines for high performance tracking. In: Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 2, pp. 1508–1515 (October 2005)
10. Rosten, E., Drummond, T.: Machine Learning for High-Speed Corner Detection. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 430–443. Springer, Heidelberg (2006)
11. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2564–2571 (November 2011)
12. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, vol. 2, pp. 1470–1477 (October 2003)
13. Wang, M., Ni, B., Hua, X., Chua, T.: Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Comput. Surv.* 44(4), 25:1–25:24 (2012)
14. Wang, W., Zhang, D., Zhang, Y., Li, J., Gu, X.: Robust spatial matching for object retrieval and its parallel implementation on gpu. *IEEE Transactions on Multimedia* 13(6), 1308–1318 (2011)
15. Xie, H., Gao, K., Zhang, Y., Tang, S., Li, J., Liu, Y.: Efficient feature detection and effective post-verification for large scale near-duplicate image search. *IEEE Transactions on Multimedia* 13(6), 1319–1332 (2011)
16. Zhang, S., Huang, Q., Hua, G., Jiang, S., Gao, W., Tian, Q.: Building contextual visual vocabulary for large-scale image applications. In: Proceedings of the International Conference on Multimedia, MM 2010, pp. 501–510. ACM, New York (2010)
17. Zhou, W., Li, H., Wang, M., Lu, Y., Tian, Q.: Binary sift: Towards efficient feature matching verification for image search. In: ACM ICIMCS, Wuhan, China (2012)
18. Zhou, W., Lu, Y., Li, H., Song, Y., Tian, Q.: Spatial coding for large scale partial-duplicate web image search. In: Proceedings of the International Conference on Multimedia, MM 2010, pp. 511–520. ACM, New York (2010)

Beyond Kmedoids: Sparse Model Based Medoids Algorithm for Representative Selection

Yu Wang, Sheng Tang, Feidie Liang, YaLin Zhang, and Jintao Li

Advanced Computing Research Laboratory, Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

{wangyu, ts, liangfeidie, zhangyalin, jtli}@ict.ac.cn

Abstract. We consider the problem of seeking representative subset of dataset, which can efficiently serve as the condensed view of the entire dataset. The Kmedoids algorithm is a commonly used unsupervised method, which selects center points as representatives. Those center points are mainly located in high density areas and surrounded by other data points. However, boundary points in the low density areas, which are useful for classification problem, are usually overlooked. In this paper we propose a sparse model based medoids algorithm (Smedoids) which aims to learn a special dictionary. Each column of this dictionary is a representative data point from the dataset, and each data point of the dataset can be described well by a linear combination of the columns of this dictionary. In this way, center and boundary points are all selected as representatives. Experiments evaluate the performances of our method for finding representatives of real datasets on the image and video summarization problem and the multi-class classification problem, and our method is shown to out-perform state-of-the-art in accuracy.

Keywords: representative subset, sparse model, dictionary learning.

1 Introduction

In the field of machine learning, computing vision and information retrieval, the scale of dataset grows at an ever increasing rate. Dealing with massive dataset is time- and memory- consuming. Thus being able to select a relatively small number of samples from a dataset, which can serve as a condensed view of the entire dataset, is of importance. Using those representative samples for classification and clustering algorithms can greatly reduce the memory requirement and computational time. In addition, representative samples can be available for online extension.

Kmedoids [1] is a common unsupervised method which produces representative samples. Similar to Kmeans, it assumes data points are distributed around several cluster centers. But unlike Kmeans, those cluster centers of Kmedoids are data points themselves, called medoids. Those medoids are usually located in the high density areas, ignoring the low density boundary zones. As shown in Fig.1 medoids of Kmedoids algorithm are mostly concentrated in high density areas of the distribution of the original dataset. But for classification problem, e.g. SVM, low density boundary areas deserves more concern [2].

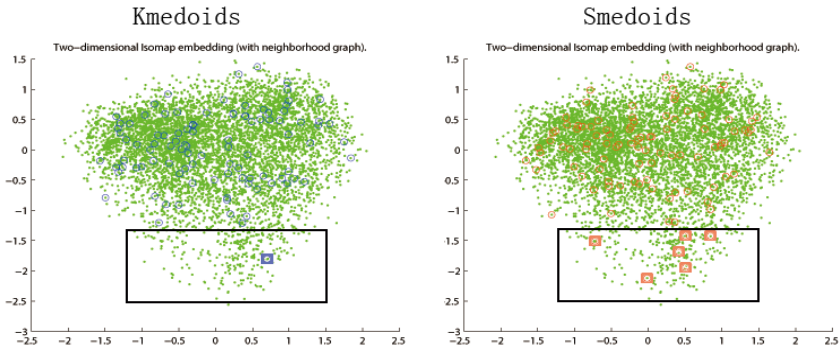


Fig. 1. Visual comparison of Smedoids and Kmedoids on digit ‘2’ of USPS. The green points represent samples of digit ‘2’. The points circled by red are representative points selected by our method, while those circled by blue are selected by Kmedoids. The representative points of Kmedoids are mostly concentrated in the high density areas of the entire distribution, e.g. there is only one representative point in the low density areas (in the black box); While ours has many representatives in the low density zones, e.g. six points are selected in the black box.

In order to get accurate condensed view of the entire dataset, we propose a sparse model based medoids algorithm (Smedoids), to find a compact dictionary whose columns are representative samples. Unlike Kmedoids, in which each data point is assigned to one center representative point, in our method each data point can be expressed by a linear combination of the representative points. In other words, representatives are subset of the entire data, which are referred most frequently by all the other data. In this way, whole distribution of original dataset can be well covered and deviation of our method between original data and representatives are less than Kmedoids’. As shown in Fig.1 representative points in our method are not necessarily the cluster centers and many are in the low density zones.

2 Related Work

This work is closely related to representative selection methods and dictionary learning methods. First, we briefly review some related representative selection methods. Then dictionary learning methods are introduced.

2.1 Representative Selection Methods

Several methods have been proposed for representative selection. Kmedoids [1] and Affinity propagation [3] are common unsupervised methods. Kmedoids, similar to Kmeans, is an iterative algorithm to find data centers surrounded by other data. When similarities between pairs of samples are given, Affinity propagation uses a message passing algorithm to find those data centers. When data are assumed to be low-rank, there are many methods [4],[5] using matrix factorization to select a few columns from the data matrix. The methods proposed by [7], [8] are recent supervised methods

for representative selection. The method by [7] aims to find representatives close to their class and far from other classes. While the method by [8] suggests a representative selection technique focused on having a large hypothesis margin to improve the performance of the 1-NN rule.

Our method is unsupervised, which is different from the above methods for the following reasons: First, representative points are not necessarily the centers as Kmedoids. In our method each data point can be described by a linear combination of representatives. Those representatives are those data points referred most frequently by other data in the dataset. Second, in the low density areas the number of representative points is more than those center methods. We illustrate that lots of points are near in high density areas. Because they are similar, they have high probability to refer the same representatives within acceptable range of error. Thus representatives in high density areas are highly reused. While in the low density areas, data points have larger distance than those high density points, so they have to refer their own neighbor points, and many representatives must be chosen in this zone. Third, our method uses dictionary learning method instead of the matrix decomposition approach, so it does not require the data to be low-rank.

Worth noting that [9] proposes a sparse modeling representative selection method (SMRS) for finding representative objects in two steps: it first uses all data points as dictionary for sparse coding and second selects representatives according to their sparse representations. But using dataset as a large redundant and coherent dictionary makes sparse coding unstable and expensive, while learning a compact dictionary can overcome those problems [11]. Our method is a dictionary learning method, which is stable and efficient. Besides we have compared with SMRS in experiments and received even better results.

2.2 Dictionary Learning Methods

Learning a dictionary from data under some constraints is widely used in computer vision and machine learning problems [6]. K-SVD algorithm [10] uses SVD decomposition of the error matrix to learn over complete dictionary from redundancy signals. Dictionaries according to many classes are constructed for clustering problem [11],[12]. Transfer learning task builds a common dictionary to find new features [13]. Task driven dictionary learning algorithm for classification is proposed by [14]. The method proposed by [15] uses online optimization based on stochastic approximation which is suitable for large-scale task.

However atoms of those dictionaries are not the original data points, hence they can't be used as representative points directly. Different from those previous works, the proposed Smedoids algorithm learns a dictionary which is subset of the dataset. Each data point in the dataset can be described as a linear combination of atoms from this learned dictionary, so this dictionary can be considered as the condensed view of the dataset.

3 Problem Formulation

Consider a set of data points in R^m arranged as the columns in data matrix $X = \{x_1, \dots, x_n\}$, the representative selection methods seek the representative matrix $D = \{d_1, \dots, d_l\}$ which are subset of X and the condense view of the original dataset.

The Kmedoids algorithm learns representative points D in two steps: first dividing data points into k parts, and each data point has only one representative in D ; Second fixing the division and finding a new medoids in each part. The problem is formulated as follows:

$$\min_a \sum_{i=1}^n \|x_i - Da_i\|_2^2, \text{ s.t. } \|a_i\|_0 = 1, \tag{1}$$

$$\min_D \sum_{i=1}^n \|x_i - Da_i\|_2^2, \tag{2}$$

Where a_i is the coefficient of x_i . The ℓ_0 -norm of a_i in Eq. (1) is to ensure that each x has only one representative point in D , and Eq. (2) solves a better D when fixing a_i . Eq. (1) and Eq. (2) can be rewritten as follow:

$$\min_{D,a} \sum_{i=1}^n \|x_i - Da_i\|_2^2, \text{ s.t. } \|a_i\|_0 = 1 \tag{3}$$

ℓ_0 -norm of a_i equals 1 constraint the solution to be center points in high density areas, but as pointed out in [2] high density areas is weak for classification. Thus in our proposed Smedoids algorithm, we extend $\|a_i\|_0 \leq s$, where s is the maximum number of nonzero items in the sparse representations, and the problem turns to be sparse modeling [16]. Since the ℓ_0 norm is NP-hard, we replace the ℓ_0 norm with the ℓ_1 norm. Different from the former works, the atoms of dictionary we get are the actual data points, which can describe the original data set. The formulation is following:

$$\min_{D,a} \sum_{i=1}^n \|x_i - Da_i\|_2^2 \quad \text{s.t. } \|a_i\|_1 \leq s, D \in X \tag{4}$$

With our method the experiments show that the distribution of the dataset can be well covered by atoms of this dictionary, and the total reconstruction error $\|x_i - Da_i\|_2^2$ is less than the Kmedoids.

3.1 Smedoids Algorithm

The Smedoids algorithm we proposed is aim to solve Eq. (4). The Eq. (4) has two variables and is not a convex problem. We can fix one variable, then the Eq. (4)

becomes a convex problem. The Smedoids performs the following two steps iteratively: first learning sparse representation of each data point using LASSO algorithm [17]; Second fixing sparse representations and finding a better D column by column. By those steps the total reconstruction error of Eq. (4) is reduced gradually as shown in Fig. 2. Details are shown in Algorithm 1.

Algorithm 1. Smedoids

Input: $X \in R^{m \times n}$, T (the number of iterations), l (the number of atoms in D), s (the maximal number of nonzero atoms in each sparse representation).

Output: $D \in R^{m \times l}$.

Initialization: ℓ_2 -normalize X , and randomly select l samples from X to initialize D.

Repeated until T

- Sparse coding: computing sparse representations $A = \{a_1, \dots, a_n\}$ using LASSO by solving:

$$\min_a \sum_{i=1}^n \|x_i - Da_i\|_2^2 \text{ s.t. } \|a_i\|_1 \leq s, D \in X \tag{5}$$

- Solving a better D: for the k -th column in D, fixing other columns:
 - ◆ Define the group of data X_{ref} , and $ref = \{i \mid 1 \leq i \leq n, a_i^k \neq 0\}$.
 - ◆ Define the sparse representation matrix A_{ref} , let $a_k = A_{ref}(k, :)$, $A_{ref}(k, :) = 0$.
 - ◆ Compute the error matrix E_k , by $E_k = X_{ref} - DA_{ref}$.
 - ◆ The total error function $g(d_k) = \|E_k - d_k a_k\|_2^2$.
 - ◆ For each $q \in ref$:

$$d_k = \min_x g(x_q) = \min_x \|E_k - x_q a_k\|_2^2 \text{ s.t. } a_k = x_q \setminus E_k. \tag{6}$$

The Smedoids algorithm is different from standard sparse modeling in the process of dictionary updating. Smedoids algorithm aims to select representatives which coincide with original data distribution. However standard sparse modeling method solves dictionary updating by ‘calculating’, which mixes the data points, so normally the dictionary is not consistent with original data distribution. In Algorithm 1 the dictionary is updated column by column. When updating the k -th column, other columns are fixed. The variable ref indicates the data points referred the k -th column, and error $g(d_k) = \|E_k - d_k a_k\|_2^2$ reflects the total error the current column causes. The next column should be the data point which can reduce g as Eq. (6). Updating corresponding a_k with d_k is suggested in [10] as an efficient implementation.

3.2 Convergence Analysis

Algorithm 1 has two steps, and in this section we prove that each step of algorithm 1 is convergent, so this algorithm is convergent.

The first step is to calculate the sparse representations of samples. For small s compared to n , the LASSO algorithm can robustly approximate the solution of Eq. (5) [10]. This assumption is natural in applications of image and video processing. Thus with fixed D this step decreases the solution of Eq. (4).

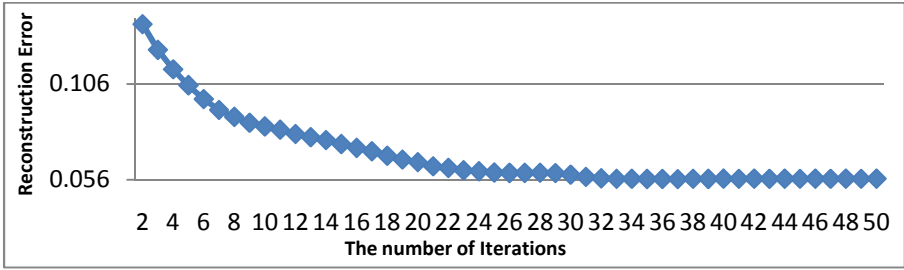


Fig. 2. Visualization of the convergence of the algorithm 1, the reconstruction error of Eq. (4) reduces gradually within 50 iterations

The second step is to find a better dictionary, when reduction or no change of the solution in Eq. (6) and the sparsity constraint are guaranteed, this step is convergent. Because the corresponding data is selected at first, the sparsity constraint is met [10]. The following proves that reduction or no change of Eq. (6)’s solution is guaranteed.

Theorem 3.1. Algorithm 1 can reduce or not change the error of the previous iteration for optimizing dictionary.

Proof: Finding a better dictionary can be rewritten as following:

$$f(D) = \min_D \|X - DA\|_2^2, \quad \text{s.t. } D \in X \tag{7}$$

Since this convex optimization problem admits separable constraints in the columns [15], updating the columns of dictionary one by one guarantees the convergence to a global optimum. When updating k -th column, the problem is following:

$$\begin{aligned} f(d_k) &= \min \|X - DA\|_2^2 = \min \|X_{nref} - DA_{nref} + X_{ref} - DA_{ref} - d_k a_k\|_2^2 \\ &= \min \|X_{nref} - DA_{nref} + E_k - d_k a_k\| \end{aligned} \tag{8}$$

where ref indicates those samples which refer the k -th column, $nref$ indicates those don’t. We set the k -th row of the matrix to zero: $A_{ref}(k, :) = 0$. In the above equation, the $(X_{nref} - DA_{nref})$ is constant since other columns are fixed, so $g(d_k) = \|E_k - d_k a_k\|_2^2$ is the total error caused by the k -th column.

For $d_k \in X$, there is a data point x_j , where $d_k = x_j$ and $x_j \in X_{ref}$. If no other data points minimize g , x_j is still selected by LASSO algorithm and the solution does not change; otherwise, other better data point is chosen to reduce the error.

From the above proof, we can conclude that algorithm 1 is convergence, so with finite iterative number T the optimized dictionary is always solved. In each iterative the complex of Lasso algorithm is $O(lmsn + ls^2n)$, and the dictionary updating is $O(ln_{ref})$. In fact on many-core platform [18], the parallelization of dictionary updating reduces the complexity to $O(n_{ref})$. Thus the total complexity of the algorithm is $T * (O(lmsn + ls^2n) + O(ln_{ref}))$. When $n > 2Tls, m \gg s$, the upper bound of this complexity is $O(mn^2)$.



Fig. 3. Visualization of the dictionary after updating with new added data, the first dictionary (on the left) is training on some samples of digit ‘1’. Then this dictionary is used as an initialization dictionary for new adding samples. The second dictionary (on the right) has learned some new atoms marked in red while other atoms are almost consistent with the first one.

3.3 Online Extension

Since the representative points coincide with original data, they can be reused when new samples in the same or similar classes are added. Let D be the representative points solved in the current dataset and X_{new} be the new added data. There are two extension methods. Let D be the initialization of the new dictionary, e.g. as shown in Fig. 3 or combine D and X_{new} into a new data collection.

When adding a new representative dictionary D_{new} of other class, since D_{new} is sufficient to describe the data collection of the other class, the optimization dictionary can be achieved simply by combining $[D D_{new}]$.

4 Experiments

In this section, we illustrate image and video summarization and multi-class classification problem for evaluating the performances of our method for finding representatives of real datasets.

4.1 Image and Video Summarization

We demonstrate that those representatives selected by the proposed algorithm can well summarize image and video datasets, so this algorithm can be used as a preprocess step in applications [24, 25].

First we consider the summarization of images of USPS dataset. The dataset consists of different variation of ten digit characters. The representatives of USPS are shown in Fig. 4 and it is worth notice that some marked representatives are not the center samples. Those marked samples have fewer occurrences than others and are hard to be classified, e.g. the marked representative of digit 4 and digit 9, digit 1 and digit 2. Those boundary representatives are useful for marginal decision in SVM training process.



Fig. 4. Representatives selected by our algorithm for the images of USPS dataset. Those representatives stand for different variation of each digit.



Fig. 5. Frames selected by our method for a one-shot video. Nine representative frames selected by our algorithm summary the activities of the video as follows: (1) a man stands by a window;(2) a man talks to someone across the widow;(3) a woman and a man enter the room;(4) one man leaves the room;(5) the first man sitting with the woman takes away her crown and hands out the window;(6) the man leaves the room; (7) the woman see the thief outside the window taking her crown;(8) the thief runs; (9) the woman passes out on the sofa.



Fig. 6. Frames selected by our method for a multi-shot video. Different numbers of representative frames are automatically computed according the amount activities of each shot as: one representative frame for the 2-th shot; three representatives for the 3-th shot; two representatives for the 4-th shot; two representatives for the 5-th shot.

Next we chose a 1,536-frame one-shot video and a 782-frame multi-shot from [19]. The one-shot video contains continuous actives in a fixed background. We use algorithm 1 to extract nine representative frames. As shown in Fig. 5 those representatives well cover the whole activities of this video. This result is better than [9] which have missed the frame where a man is passing the crown to the window. The 5-shot video contains 4 different scenes, including cartoon, sky, virtual scene and rocket launch. We extract eight representative frames. Those representatives capture all those scenes but not all those shots as shown in Fig. 6. It's noteworthy that 1-th shot is similar with 5-th shot regardless background or activity, so they share representatives, that's why 1-th shot has no representatives. Different number of representatives reflects the number of activities in each shot. Two kids are throwing a ball in the third shot. Three representatives capture the main activities of this event: (1) a boy is ready to throw a ball; (2) a girl receives the ball; (3) the girl holds the ball. Relatively static shot has fewer representatives as 2-th shot.

4.2 Classification Performance Using Representatives

We now evaluate the performance of our method as well as other algorithms for finding representatives that are used for multi-class classification problem. For training set in each class we only select a few representatives and use them as a reduced training dataset. It is believed that the better those representatives condense the original training data, the higher accuracy the classification results would get.

We compare the proposed algorithm, Smedoids, with several state-of-the-art methods for finding representatives: Kmedoids, Sparse Modeling Representative Selection (SMRS) and simple random selection (Rand). Two standard classification algorithms, multi-class classifier (SVM) [20] and Sparse Representation-based Classification (SRC) [23], are used to evaluate the multi-class classification performance. The experiments are run on the handwritten digits database USPS [21] and the Yale Face Database B [22]. The USPS handwritten database contains 11000 images of ten digit characters, and in each class 1000 samples are randomly selected for training and left for testing. The Yale-B contains 5760 images of 10 subjects which have been cropped to the size of 16 by 19 pixels by us, and in each class 300 samples are randomly selected for training and left for testing. We run several times with different number l of representatives selected from training set. Obviously with

more training representatives, the classification will achieve higher accuracy. Tables 1 and 2 show the classification results for the USPS database and the Yale-B database respectively.

From the results, we can conclude that our proposed method always gets the best accuracy. All representative selection methods work better with SVM than with SRC. In contrast to [9] SMRS performs better than Kmedoids in some cases but not always.

Table 1. Classification results on USPS digit dataset using l representatives of the 1000 training samples in each class

USPS		Rand	Kmedoids	SMRS	Proposed
Representatives #					
SRC	$l=10$	0.77	0.838	0.824	0.86
	$l=20$	0.86	0.872	0.868	0.896
	$l=30$	0.895	0.898	0.902	0.92
	$l=40$	0.9127	0.917	0.917	0.928
SVM	$l=10$	0.8027	0.8758	0.8556	0.907
	$l=20$	0.8809	0.888	0.8697	0.9236
	$l=30$	0.9137	0.9308	0.9243	0.9464
	$l=40$	0.9346	0.9392	0.9305	0.9489

Table 2. Classification results on Yale-B Face dataset using l representatives of the 300 training samples in each class

Yale-B Face		Rand	Kmedoids	SMRS	Proposed
Representatives #					
SRC	$l=10$	0.44	0.5117	0.4633	0.54
	$l=20$	0.57	0.6397	0.5947	0.6877
	$l=30$	0.6347	0.7067	0.6990	0.74
	$l=40$	0.7003	0.7627	0.754	0.7843
SVM	$l=10$	0.5873	0.6417	0.6017	0.7083
	$l=20$	0.7213	0.7877	0.7787	0.834
	$l=30$	0.7997	0.8277	0.8397	0.868
	$l=40$	0.8443	0.889	0.8787	0.9003

Our proposed method works best because not only center points but also boundary points are selected. We investigate the effect of the parameters of algorithm1, T (the number of iterations) and s (the maximal number of nonzero atoms in each sparse representation). We set T to 20 and s to 5 in all runs, and we also constraint the sparse representations to be non-negative to get the reported result.

5 Conclusion

In this paper, we propose a Smedoids algorithm to select representatives from entire dataset and prove its convergence. The Smedoids algorithm selects both center and boundary points that well cover the whole distribution of dataset, and we argue that our method can condense the original dataset better than the state-of-the-art representative selection methods. Results of video summarization show that main activities of each shot are well captured. In addition our proposed method always achieves the best accuracy among the state-of-art algorithms using representatives for multi-class classification problem.

Acknowledgement. This work was supported in part by the National Nature Science Foundation of China (61173054, 61271428); and Co-building Program of Beijing Municipal Education Commission.

References

1. Kaufman, L., Rousseeuw, P.: Clustering by means of medoids. In: Dodge, Y. (ed.) *Statistical Data Analysis based on L1 Norm*. North-Holland, Amsterdam (1987)
2. Jurie, F., Triggs, B.: Creating Efficient Codebooks for Visual Recognition. In: *ICCV* (2005)
3. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. *Science* (2007)
4. Boutsidis, C., Mahoney, M.W., Drineas, P.: An improved approximation algorithm for the column subset selection problem. In: *Proc. SODA* (2009)
5. Balzano, L., Nowak, R., Bajwa, W.: Column subset selection with missing data. In: *NIPS Workshop on Low-Rank Methods for Large-Scale Machine Learning* (2010)
6. Tang, S., Zheng, Y.-T., Wang, Y., Chua, T.-S.: Sparse Ensemble Learning for Concept Detection. *IEEE Trans on Multimedia* 14(1), 43–54 (2012)
7. Bien, J., Tibshirani, R.: Prototype selection for interpretable classification. *The Annals of Applied Statistics* (2011)
8. Marchiori, E.: Class conditional nearest neighbor for large margin instance selection. *IEEE Trans. PAMI* 32(2), 364–370 (2010)
9. Elhamifar, E., Sapiro, G., Vidal, R.: See all by looking at a few: sparse modeling for finding representative objects. In: *CVPR* (2012)
10. Aharon, M., Elad, M., Bruckstein, A.M.: The k-svd: An algorithm for designing of overcomplete dictionaries for sparse representations. *IEEE Trans. SP* 54(11), 4311–4322 (2006)
11. Ramirez, P., Sprechmann, G.: Classification and Clustering via Dictionary Learning with Structured Incoherence and Shared Features. In: *CVPR* (2010)
12. Sprechmann, P., Sapiro, G.: Dictionary Learning and Sparse Coding for Unsupervised Clustering. In: *ICASSP* (2010)
13. Raina, R., Battle, A., Lee, H., Packer, B., Ng, A.Y.: Self-taught learning: transfer learning from unlabeled data. In: *ICML* (2007)
14. Mairal, J., Bach, F., Ponce, J.: Task-driven dictionary learning. *IEEE Trans. on PAMI* 34(4), 791–804 (2011)
15. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research* 11, 19–609 (2010)

16. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. *IEEE Trans. on PAMI* 31(2), 210–227 (2009)
17. Tibshirani, R.: Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B* 58(1), 267–288 (1996)
18. Zhang, Y., Yan, C., Dai, F., Ma, Y.: Efficient Parallel Framework for H.264/AVC Deblocking Filter on Many-core Platform. *IEEE Trans. on Multimedia* 14(3), 510–524 (2012)
19. Vidal, R.: Recursive identification of switched ARX systems. *Automachine* (2008)
20. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology* 3(2), 1–27 (2011), Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
21. Hull, J.: A database for handwritten text recognition research. *IEEE TPAMI* (1994)
22. Lee, K.C., Ho, J., Kriegman, D.: Acquiring linear subspaces for face recognition under variable lighting. *IEEE TPAMI* (2005)
23. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. *IEEE Trans. on PAMI* 31(2), 210–227 (2009)
24. Wang, M., Hong, R., Li, G., Zha, Z.-J., Yan, S., Chua, T.-S.: Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification. *IEEE Trans. on Multimedia* 14(4), 975–985 (2012)
25. Hong, R., Wang, M., Xu, M., Yan, S., Chua, T.-S.: Dynamic Captioning: Video Accessibility Enhancement for Hearing Impairment. In: *ACM MM* (2010)

Improving Automatic Image Tagging Using Temporal Tag Co-occurrence^{*}

Philip McParlane, Stewart Whiting, and Joemon Jose

The University of Glasgow,
Glasgow, G12 8QQ, UK

{p.mcparlane.1,s.whiting.1}@research.gla.ac.uk, Joemon.Jose@glasgow.ac.uk

Abstract. Existing automatic image annotation (AIA) systems that depend solely on low-level image features often produce poor results, particularly when annotating real-life collections. Tag co-occurrence has been shown to improve image annotation by identifying additional keywords associated with user-provided keywords. However, existing approaches have treated tag co-occurrence as a static measure over time, thereby ignoring the temporal trends of many tags. The temporal distribution of tags, however, caused by events, seasons and memes, etc, provides a strong source of evidence beyond keywords for AIA. In this paper we propose a temporal tag co-occurrence approach to improve AIA accuracy. By segmenting collection tags into multiple co-occurrence matrices, each covering an interval of time, we are able to give precedence to tags which not only co-occur each other, but also have temporal significance. We evaluate our approach on a real-life timestamped image collection from Flickr by performing experiments over a number of temporal interval sizes. Results show statistically significant improvements to annotation accuracy compared to a non-temporal co-occurrence baseline.

Keywords: Image Annotation, Tag Co-occurrence, Temporal.

1 Introduction

With the amount of multimedia data rapidly increasing, it becomes important to organize this content effectively. To be able to facilitate efficient multimedia retrieval we must first categorize these objects with semantic features, such as keywords¹. However, unlike traditional text retrieval which can infer topics directly from the distributions of words in a document, multimedia objects provide little or no textual clues. Hence, content annotation with semantically related keywords is therefore necessary before indexing and retrieval can take place. The laborious nature of manual image annotation, however, combined with the need for effective large-scale image search has increased research in the field of automatic image annotation (AIA).

^{*} This research was supported by the the European Community's FP7 Programme under grant agreements nr 288024 (LiMoSINe).

¹ For the remainder of this paper we refer to tags and keywords synonymously.

Current state-of-the-art AIA models, however, produce poor results, especially when tested on ‘real-world’ image collections [2]. Such collections are considered problematic often because of their noisiness, sparsity and diversity of image features. Bridging the semantic gap between low-level image features and high-level human concepts is still an unsolved research problem [23]. In any case, many fundamentally question if there even exists a correlation between these two levels [21]. Much research has focused on looking *beyond the pixel* to incorporate more robust evidence in the annotation process [20,16]. We propose to explore beyond the visual contents of images in the annotation process by exploiting tag co-occurrence and temporality; by doing so we can avoid, to an extent, the problems associated with content-based image annotation.

Since the quality of AIA is very poor, a number of image sharing websites employ user tagging e.g. Flickr. However, the tagging process is either incomplete or often inaccurate. Automatic tagging techniques are often exploited to improve the quality of annotated tags. Tag co-occurrence has been used by existing tag recommendation [22] and AIA systems [16] to improve performance by discovering additional related tags. Tag co-occurrence for two keywords is defined as the number of documents in which both keywords co-exist; in the field of AIA, these documents are images. The motivation for exploiting tag co-occurrence is that keywords exist in a specific distribution which can be exploited. In the field of timestamped text analysis, a significant body of research has sought to exploit dynamic term distributions, most notably for Topic Detection & Tracking [1] and IR [26]. Analysis of user tags shows that tag co-occurrence is often linked with time. As such, two keywords which co-occur highly in June may not have the same relationship in December. Figure 1 shows example normalised tag distributions over time from a collated Flickr collection. Strong temporal distributions are seen for seasonal keywords such as **summer** and **winter**, which is expected. Further, tags related to weather cycles also observe a relationship with time. For example, **frost** and **snow** are most prominent during the winter months.

It may be argued, however, that only a restricted set of seasonal and weather related keywords will display such strong temporal distributions in image annotation but actually there are many tags with implicit temporality. For example, keywords such as **jet** and **pool** are seen to increase during the summer months i.e. typically when people go on vacation. Similarly, **garden** observes peaks during May through September which is expected due to the increase in outdoor activities in summer. By harnessing these temporal trends, we propose to improve tagging accuracy of an existing state-of-the-art model. Finally, research into tag co-occurrence has implications for a number of fields such as: tag recommendation systems as used on social bookmarking websites [5], query expansion [10], event detection [27] and personalised IR [3].

This paper is organised as follows. In Section 2 we present related work in the field of automatic image annotation and temporal IR. Section 3 describes the methodology behind our temporal co-occurrence based approach. In Section 4 we discuss our experimental setup. Finally, Section 5 presents the results of our experiments and Section 6 concludes and discusses avenues for future work.

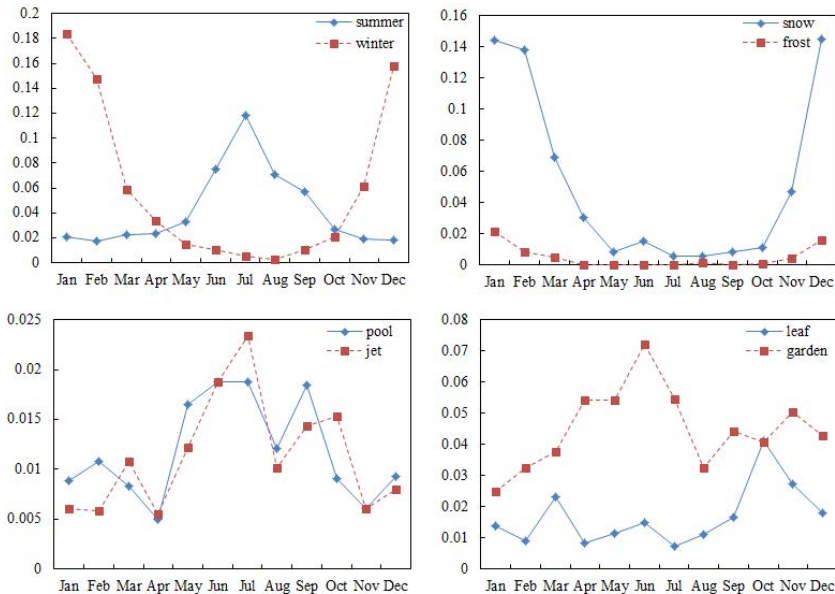


Fig. 1. Tag distributions over time from our Flickr Collection

2 Related Work

The problem of image classification is often treated as a cross media modelling problem where we try to map low-level features in vector format to high-level textual concepts. Duygulu *et al.* [6] treated the problem of image annotation using a machine translation approach where images are segmented into small regions; keywords were then mapped based on a number of image features. In 2003, Joen *et al.* [9] adopted the cross lingual language model of Lavrenko *et al.* [13], Cross-Media Relevance Models (CMRM), to predict the probability of generating a word given blobs in an image in the training set. The model assumed regions in an image can be described by a small vocabulary of blobs, which were created from image features using clustering techniques. Lavrenko *et al.* [14] then proposed the Continuous-space Relevance Model (CRM) which generalised the previous CMRM to model highly dimensional continuous features without clustering and quantization. Bag of Visual Words (BOVW) has gained much interest in the field; Carneiro *et al.* [4] proposed a Gaussian mixture model using the bag of local features approach for class conditional dependencies.

More recently, Makadia *et al.* showed that all of the previously stated models could be outperformed by adopting a K-nearest neighbour approach trained on Gabor and HAAR image features [18]. In a similar experiment, Athanasakos *et al.* showed that these approaches were also out-performed by using an SVM approach trained on global features [2]. Further, they highlighted problems of the evaluation approaches of state-of-the-art annotation (SOTA) models, which

are addressed in Section 4.2. We have chosen to implement the approach by Athanasakos *et al.* as a baseline, due to its simplicity and performance against other SOTAs.

Following research in text based IR [5,10,27,3], tag co-occurrence has been used as a secondary source of evidence in tag recommendation systems [22] and image annotation models [17,16]. Sigurbjornsson *et al.* proposed a tag recommendation strategy to support users annotating photos on Flickr [22]. The relationships between tags were exploited to suggest highly co-occurring tags. Sigurbjornsson *et al.* adopted two normalised measures for tag co-occurrence: the Jaccard (symmetric) and Asymmetric coefficients. Our approach follows this research by using these coefficients as a measure of keyword similarity. Llorente *et al.* incorporated tag co-occurrence in their annotation model which formulated the problem of image annotation as that of direct image retrieval [17]. Novelty was achieved by not only exploring the dependencies between words and their semantic context, but also between visual features and words.

Temporality has previously been studied and exploited in both information seeking and retrieval systems. Despite this, its implication on automatic image annotation has not yet been explored. Kleinberg *et al.* [12] developed a framework for modelling periodic bursts of keywords in a corpus with hierarchical structure using an infinite-state automaton. More recently, Leskovec *et al.* [15] performed a large-scale study of “memes” diffusing throughout news media as a result of temporal rhythms. As a result, a mathematical model was provided for analysing the temporal variation in the context of news. We propose to exploit these temporal trends of tags in a tag co-occurrence model.

3 Temporal Co-occurrence

In this section we present our temporal based co-occurrence approach for improving the effectiveness of tag suggestions made by an existing AIA model.

3.1 Problem Statement

Let $I = \{i_1, \dots, i_m\}$ denote an image collection, where m is the number of images in the image set. We denote t as a tag and $T = \{t_1, \dots, t_n\}$ our vocabulary, where n is the number of keywords in our collection. We define $S(i_x, t_y)$ as a confidence score of matching tag t_y to image i_x .

Every $i \in I$ has a time-stamp of when it was taken. We aim to cluster images based on time. We therefore define β to be the number of time intervals in the year in which we wish to cluster images on. For example, $\beta = 3$ would group images into three, $122 \left(\frac{366}{3}\right)$ day, time intervals. We define $i_z \subset I$ where i_z is a set of images taken between the start and end of time interval z , where $1 \leq z \leq \beta$.

Our approach improves image annotation by promoting the most highly co-occurring tags from our image classifier. For each subset of images taken within a given time interval, $i_z \in I$, we build a co-occurrence matrix C_z mapping the number of images two tags co-occur in for the given time interval.

$$C_3 = \begin{matrix} & t_1 & t_2 & \dots & t_n \\ \begin{matrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{matrix} & \begin{bmatrix} 0 & 10 & \dots & 1 \\ 10 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \dots & 0 \end{bmatrix} \end{matrix} \tag{1}$$

where C_3 is the matrix constructed from images taken within the 3rd interval *e.g.* tag t_1 occurs together with tag t_2 in 10 images. Tag co-occurrence measures, however, are actually normalised between 0 and 1, as explained in Section 3.3. We define $C_{overall}$ to be the co-occurrence matrix built from all images.

3.2 Content Based Annotation

Our proposed approach builds on top of a linear SVM based AIA approach. SVMs have been used for many years in text based information retrieval categorisation systems [24]. More recently, this methodology has been used in AIA systems and has been seen to outperform state-of-the-art annotation models [2]. Due to its performance against other baselines and simplicity in design, we will use this model as our baseline to improve upon.

We implement the SVM_{light} model [11], which uses a linear kernel function, trained upon the MPEG-7 Global Edge Histogram (GEH) image feature [19]. This feature was seen to give greatest annotation accuracy in [2]. Our approach is to train n classifiers in an one-versus-all scheme, where n is the number of classes (tags). We use the normalised distance $-1 \leq d_{xy} \leq 1$ to the boundary plane as a measure of how trustworthy a tag t_x is for a given image i_y . Therefore, we define $S(i_x, t_y) = d_{xy}$.

It could be argued that our approach could be improved by training $n(n-1)/2$ classifiers in the one-versus-one scheme where we train a SVM for every tag and every tag *combination*, thus retaining prior classification data. We argue, however, that this would quickly become computationally challenging as n increases. For example, for our collection containing 270 tags, we would potentially have to train 36,315 SVMs. In a real-world collection where there are millions of keywords [22], this solution would become unscalable. Further, this method requires a heavily dense collection with each tag combination containing sufficient training data, which is not true in on-line collections [25].

3.3 Improving Annotation through Tag Co-occurrence

To improve annotations made by the SVM, we increase or decrease $S(i_x, t_y)$, using tag co-occurrence measures. $S(i_x, t_y)$ is therefore redefined as:

$$S(i_x, t_y) = \lambda \cdot P_{svm}(i_x, t_y) + (1 - \lambda) \cdot P_{cooc}(i_x, t_y) \tag{2}$$

where $P_{svm} = d_{xy}$. P_{cooc} is the tag co-occurrence score for t_y with the other SVM suggested tags. λ is a parameter ($0 \leq \lambda \leq 1$) which weights the amount of SVM and co-occurrence data we use for $S(i_x, t_y)$. $P_{cooc}(i_x, t_y)$ is as follows:

$$P_{coc}(i_x, t_y) = \frac{\sum_{t_w \in T_{svm} - t_y} C(t_w, t_y)}{|T_{svm}|} \quad (3)$$

where T_{svm} is the set of tags suggested by the SVM (where $d \geq 0$), $|T_{svm}|$ is the number of tags suggested by SVM and $C(t_w, t_y)$ is the tag co-occurrence frequency between tag t_w and t_y . Effectively, keywords in the SVM prediction set are promoted if they co-occur highly with the rest of the predictions, and demoted otherwise.

Baseline. Our tag co-occurrence baseline takes normalised tag co-occurrence frequencies $C(t_w, t_y)$ from $C_{overall}$. Therefore, co-occurrence frequencies are static and taken from the entire collection, thus ignoring temporality.

Temporal. Our temporal approach takes co-occurrence frequencies from the temporal interval in which the image was taken. e.g. if $\beta = 12$ (equivalent to 1 matrix per month) and an image is taken on the 15th of March, co-occurrence scores are taken from C_3 .

Using raw tag co-occurrence frequency is noisy, however, as the popularity of tags is not taken into account. This gives rise to weighting popular tags higher than less common keywords; we must first normalise these frequencies. We have decided to use two measures as chosen by previous work, namely the Jaccard and Asymmetric Measures: [22]:

$$J(t_i, t_j) = \frac{|t_i \cap t_j|}{|t_i \cup t_j|} \quad P(t_j|t_i) = \frac{|t_i \cap t_j|}{|t_i|} \quad (4)$$

Equation 4. The Jaccard (left) and Asymmetric (right)

Both measures which are used to compute tag similarity and relatedness have an upper bound of 1 and a lower bound of 0. Previous work has stated that the Jaccard measure is more useful for identifying synonyms whereas the Asymmetric measure offers more diverse recommendations. We will compare the effectiveness of both measures in our work.

4 Experiments

Our experiments compare annotation accuracy made by three systems:

- **SVM (Contents)** The first system is the state-of-the-art (as defined in [2]) which annotates using SVM data only.
- **SVM^{Coc} (Contents + Co-occurrence)** Our baseline improves results from the SOTA by exploiting tag co-occurrence data.
- **SVM^{TempCo} (Contents + Temporal Co-occurrence)** Our experimental approach improves on SVM^{Coc} by exploiting temporal information in the computation of tag co-occurrence measures.

4.1 Collection

We tested our approach on a collated real life image collection from Flickr². Real life image collections pose problems for research as the tags are *inconsistent*, often *misspelt* and *sparse* (many tags are used in only one image). We therefore cleaned the collection to contain only tags which occurred in at least 40 images and where images contained at least 3 tags. We also filtered out tags which, when classified by WordNet [7], were not considered *nouns*. This would remove tags which were not suitable for AIA; for example, subjective (e.g. *so cute*, *nice*) and organisational tags (e.g. *me*, *avoid*). Once cleaned, the collection contained 12,985 images and 270 tags. Each image on average contained 4.07 tags.

4.2 Experimental Procedure and Settings

Our experiments are taken out in a two stage process. Initially, images are trained and tested on the keywords using a linear SVM based on the Global Edge Histogram feature, as described in Section 3.2. For each image, a list of keyword scores is returned, measuring the likelihood of a tag occurring in an image. Tag co-occurrence is then employed as a reweighing scheme by increasing or decreasing the given score for a tag, based on its co-occurrence with the other tags in the ground truth. After this reweighing stage, the tags with the highest scores are selected for annotation; the amount of tags selected is equal to that of the number of tags in the image's ground truth.

We introduce temporality by computing the tag co-occurrence measures in predefined intervals, whereas our baseline computes tag co-occurrence measures over the entire year i.e. 1 co-occurrence matrix. We varied our temporal interval size over a range of values from half a year to 2 days. Therefore, given a new image i_x , we select the co-occurrence measures from the co-occurrence matrix which is built upon images taken *in the same time interval* as image i_x . We compare results when using the following number of co-occurrence matrices: $\beta = 2, 6, 12, 18, 40, 52, 70, 90, 120$. For each interval size, we compare annotation accuracy between our 3 approaches using 10-fold cross validation over 10 iterations. For each iteration of the experiment we collate a *subset* of the overall collection which is *smoothed* and *normalised*. By *smoothed* and *normalised* we mean that most of the tags in the test collection contain approximately the same number of training images. Alternatively, popular keywords, such as *sky*, and unpopular keywords, such as *hammer*, are *not* selected for testing.

We have taken out this stage as using the whole collection would probably create an easier evaluation setting due to the following reasons. Firstly, by evaluating on the entire collection, popular keywords would more likely be selected for testing. Secondly, when annotating an image, the model would be more likely to select a more frequent tag. By normalising our collection we create a fairer evaluation test-bed where images are less likely to be annotated with tags based purely on their popularity. We therefore normalise our collection as is explained

² <http://www.flickr.com/>

in [2]. This stage is an important stage in our experiment, as it will reduce the perceived accuracy of our state-of-the-art, as the test collection is more “difficult” as the number of perceived “easy” keywords is reduced in the test collection.

For each iteration of the experiment, on average a subset of 1114 images were used for training and 124 used for testing. We tested using 100 keywords at each iteration, with each keyword containing at least 20 training images. In our experiments we compute *precision*, *recall* and the *number of words recalled*.

5 Results and Discussion

The following sections detail the results of our experiments showing the potential of temporal modelling in the annotation process and the conditions where accuracy is maximised.

5.1 Effects of Temporality

By exploiting temporality, we were able to achieve statistically significant improvements to annotation accuracy. Table 1 shows the results of our experiments comparing both normalization methods over all interval sizes. From Table 1 we can see that both coefficients give increases to recall, precision and number of words recalled when compared to our static baseline. Using the Jaccard coefficient produces marginally better results than when using the asymmetric co-occurrence measure. It may be noted that the measures appear somewhat low for a SOTA; this is a side effect of the collection normalisation as described in Section 4.2. In effect our model is annotating on a *difficult* subset of an already *difficult* real-life image collection, hence lower performance is expected.

Figure 2 illustrates the conditions where annotation accuracy is maximised. The scores in Figure 2 are taken as an average of recall, precision and number of words recalled over the baseline. The Jaccard coefficient produces best results when the interval window size is set to 6 days ($\beta = 70$); statistically significant improvements averaging 11.6% are observed. The Asymmetric coefficient produces best results using approximately the same interval size, i.e. 5 days ($\beta = 90$), achieving an 9.6% increase to annotation accuracy. By incorporating the temporal trends of tags in images, as seen in Figure 1, we are able to give precedence to temporally significant tags based on the time an image is taken, thus improving the annotation accuracy.

Using a large interval size, 183 days for example, has a slight detrimental effect on AIA accuracy however. This may be because temporal profiling barely exists at these levels. We believe it may have the opposite effect of adding noise to the co-occurrence measures. We therefore recommend that future temporal profiling of keywords should use a interval size of around 5 days.

Finally, λ was trained giving a local maxima in annotation performance when $\lambda = 0.4$. Interestingly we achieve greatest accuracy when we use a higher weight of P_{cooc} than P_{svm} implying tag co-occurrence and temporality may be a more reliable source than image contents in the annotation process.

Table 1. Each column denotes the scores for each measure using the given number of intervals. On the left column, R, P and W stand for recall, precision and the number of words recalled respectively. Bolded columns denote the interval size which produced the largest average improvement over the baseline. Paired t-test statistical significance comparing our experimental approach against the baseline are denoted as * being $p < 0.05$, ** being $p < 0.01$ and *** being $p < 0.001$.

Using the Jaccard Co-efficient										
Intervals	2	6	12	18	40	52	70	90	120	
R	SVM	0.0809	0.0782	0.0796	0.0769	0.0778	0.0732	0.0807	0.0726	0.0745
	SVM ^{C_{ooc}}	0.0839	0.0789	0.0825	0.0824	0.0778	0.0755	0.0814	0.0757	0.0782
	SVM ^{T_{empCo}}	0.0842	0.0816	0.0853	0.0836	0.0838**	0.0813**	0.0909***	0.0842**	0.0855*
P	SVM	0.0697	0.0619	0.0712	0.0598	0.0759	0.0596	0.0637	0.0691	0.0619
	SVM ^{C_{ooc}}	0.0719	0.0588	0.0750	0.0637	0.0720	0.0610	0.0613	0.0709	0.0632
	SVM ^{T_{empCo}}	0.0717	0.0625	0.0738	0.0640	0.0742	0.0649	0.0695***	0.0784*	0.0718*
W	SVM	20	19.5	20.3	18.9	20	18.4	18.1	19.3	18.5
	SVM ^{C_{ooc}}	20.4	19.4	21	19.7	19.9	18.7	18.2	19.7	19.1
	SVM ^{T_{empCo}}	20.2	19.8	21.3	20.3	21.1**	20.3***	20.1***	21.5**	20.7*
+/- Over Baseline		-0.3%	+3.9%	+1.0%	+1.7%	+5.6%	+7.6%	+11.9%	+10.3%	+10.5%
Using the Asymmetric Co-efficient										
Intervals	2	6	12	18	40	52	70	90	120	
R	SVM	0.0809	0.0782	0.0796	0.0769	0.0778	0.0732	0.0807	0.0726	0.0745
	SVM ^{C_{ooc}}	0.0849	0.0824	0.0855	0.0823	0.0786	0.0778	0.0862	0.0787	0.0795
	SVM ^{T_{empCo}}	0.0837*	0.0848	0.0886	0.0862	0.0834**	0.0843**	0.0909*	0.0873**	0.0851*
P	SVM	0.0697	0.0619	0.0712	0.0598	0.0759	0.0596	0.0637	0.0691	0.0619
	SVM ^{C_{ooc}}	0.0720	0.0628	0.0724	0.0630	0.0689	0.0589	0.0639	0.0722	0.0615
	SVM ^{T_{empCo}}	0.0711*	0.0628	0.0742	0.0662	0.0677	0.0644**	0.0714**	0.0786*	0.0676
W	SVM	20	19.5	20.3	18.9	20	18.4	18.1	19.3	18.5
	SVM ^{C_{ooc}}	20.4	19.8	21.2	19.8	19.7	18.9	18.6	19.9	19.1
	SVM ^{T_{empCo}}	20*	19.9	21.5	20.4	20.7**	20.5**	20.5***	21.7**	20.1
+/- Over Baseline		-1.6%	+1.1%	+2.5%	+4.3%	+3.1%	+8.7%	+9.1%	+9.6%	+7.4%

5.2 Tag Distributions

The following section gives real life examples of tags with high temporal relationships. Figure 3 shows the co-occurrence frequencies of temporarily significant keywords, **snow** and **winter**, with **tree** and **landscape** over 18 time intervals.

We can clearly observe the keywords’ correlation with time. Both sets of keywords co-occur highly at the beginning and end of the year with almost no co-occurrence during time intervals 5 through 16. This produces different Jaccard and

Table 2. Jaccard and Asymmetric scores

Measure	Scores @ Interval		
Time Interval	All	2	10
J(tree, snow)	0.09	0.10	0
A(tree, snow)	0.18	0.28	0
J(landscape, winter)	0.05	0.08	0
A(landscape, winter)	0.09	0.22	0

Asymmetric measures at different periods in the year. Table 2 compares these co-occurrence measures at different time intervals.

The temporal distribution shown in Figure 3 highlights that keyword co-occurrence measures should consider time in these calculations. In our example, **snow** only exists along side images of **trees** in images during the winter months.

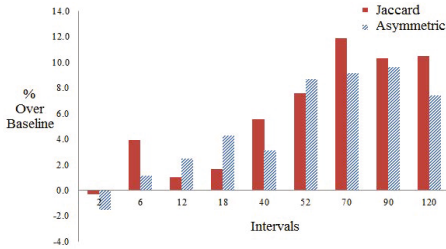


Fig. 2. Co-occurrence measures

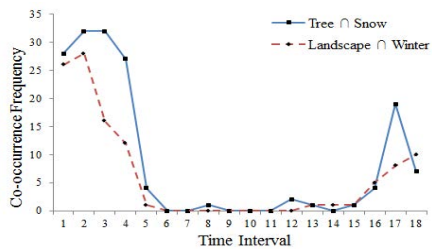


Fig. 3. Tag Co-occurrence distributions

Our baseline which ignores temporal profiling of tag co-occurrences can be represented by the *entire* column of Table 2. Columns 2 and 10 show the coefficient scores for the given time intervals only. These intervals were chosen as they are the most divergent coefficient scores for the given keywords over the year. The coefficients in time *interval 2* are 73% higher, on average, than those taken over the whole year. We believe this is logical as the keywords, *tree* with *snow* and *landscape* with *winter*, co-occur most frequently in this time interval. Similarly, all the coefficients compute as 0 in *interval 10* as the keywords never co-occur in this time period. We believe this is sensible: if two keywords never co-occur in a given time period, they should produce a co-occurrence coefficient of 0 regardless of if they co-occur in other time intervals. By ignoring this noise and placing higher precedence to *temporally significant* keywords, we are able to achieve improvements to AIA accuracy.

6 Conclusion and Future Work

Accurate automatic image annotation is highly desired to be able to build effective multimedia retrieval systems. In this work we present a novel temporal based tag co-occurrence technique for the improvement of a state-of-the-art SVM based automatic image annotation model. Results from our experiments show that by exploiting temporal tag co-occurrences, we can produce statistically significant improvements to AIA accuracy.

This paper argues that static measures of normalised tag co-occurrence as used by previous methods are insufficient and that keywords co-occur in a non linear temporal distribution which can be exploited. We achieve this by constructing a number of co-occurrence matrices, one for each predefined interval, instead of building a co-occurrence matrix over the entire year. We further experiment by changing the interval sizes used to construct the co-occurrence matrices. We conclude that best results are achieved when the size of the temporal window is set to 5 or 6 days. Future work will look at extending our exploitation of temporal tag co-occurrence for AIA by incorporating more sophisticated techniques from temporal text-based IR systems.

References

1. Allan, J.: Topic detection and tracking, pp. 1–16. Kluwer Academic Publishers, Norwell (2002)
2. Athanasakos, K., Stathopoulos, V., Jose, J.M.: A framework for evaluating automatic image annotation algorithms. In: ECIR 2010 Milton Keynes, UK (2010)
3. Byde, A., Cayzer, S.: Personalized tag recommendations via tagging and content-based similarity metrics, New York (2) (2007)
4. Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions* 29 (2007)
5. Cattuto, C., Benz, D., Hotho, A., Stumme, G.: Semantic Grounding of Tag Relatedness in Social Bookmarking Systems. In: Sheth, A.P., Staab, S., Dean, M., Paolucci, M., Maynard, D., Finin, T., Thirunarayan, K. (eds.) ISWC 2008. LNCS, vol. 5318, pp. 615–631. Springer, Heidelberg (2008)
6. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
7. Fellbaum, C. (ed.): *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, illustrated edn. The MIT Press (1998)
8. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin del la Société Vaudoise des Sciences Naturelles* (1901)
9. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: SIGIR 2003, NY, USA, pp. 119–126 (2003)
10. Jin, S., Lin, H., Su, S.: Query expansion based on folksonomy tag co-occurrence analysis, pp. 300–305 (August 2009)
11. Joachims, T.: *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, Norwell (2002)
12. Kleinberg, J.: Bursty and hierarchical structure in streams. In: KDD 2002, pp. 91–101. ACM, New York (2002)
13. Lavrenko, V., Choquette, M., Croft, W.B.: Cross-lingual relevance models. In: SIGIR 2002, pp. 175–182. ACM, New York (2002)
14. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: NIPS. MIT Press (2003)
15. Leskovec, J., Backstrom, L., Kleinberg, J.: Meme-tracking and the dynamics of the news cycle. In: KDD 2009, pp. 497–506. ACM, New York (2009)
16. Li, W., Sun, M.: Automatic Image Annotation Based on WordNet and Hierarchical Ensembles. In: Gelbukh, A. (ed.) CICLing 2006. LNCS, vol. 3878, pp. 417–428. Springer, Heidelberg (2006)
17. Llorente, A., Manmatha, R., Rüger, S.: Image retrieval using markov random fields and global image features. In: CIVR 2010, pp. 243–250. ACM, NY (2010)
18. Makadia, A., Pavlovic, V., Kumar, S.: Baselines for image annotation. *Int. J. Comput. Vision* 90(1), 88–105 (2010)
19. Manjunath, B.S.: *Introduction to MPEG-7, Multimedia Content Description Interface*. John Wiley and Sons, Ltd. (2002)
20. Monaghan, F., O’Sullivan, D.: Leveraging Ontologies, Context and Social Networks to Automate Photo Annotation. In: Falcidieno, B., Spagnuolo, M., Avrithis, Y., Kompatsiaris, I., Buitelaar, P. (eds.) SAMT 2007. LNCS, vol. 4816, pp. 252–255. Springer, Heidelberg (2007)

21. Santini, S., Gupta, A., Jain, R.: Emergent semantics through interaction in image databases 13(3), 337–351 (2001)
22. Sigurbjörnsson, B., van Zwol, R.: Flickr tag recommendation based on collective knowledge. In: WWW 2008, pp. 327–336. ACM, New York (2008)
23. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years 22(12), 1349–1380 (2000)
24. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc., New York (1995)
25. Weinberger, K.Q., Slaney, M., Van Zwol, R.: Resolving tag ambiguity. In: MM 2008, pp. 111–120. ACM, New York (2008)
26. Whiting, S., Moshfeghi, Y., Jose, J.M.: Exploring term temporality for pseudo-relevance feedback. In: SIGIR 2011, pp. 1245–1246. ACM, New York (2011)
27. Yao, J., Cui, B., Huang, Y., Zhou, Y.: 2010 IEEE 26th International Conference on Data Engineering (ICDE), pp. 780–783 (March 2010)

Robust Detection and Localization of Human Action in Video

Haojie Li¹, Fuming Sun^{2,*}, and Yue Guan¹

¹ School of Software, Dalian University of Technology

² Liaoning University of Technology

hjli@dlut.edu.cn, sunfm@mail.ustc.edu.cn, worm004@hotmail.com

Abstract. We propose a robust and efficient method for accurate detecting and localizing complex human action in video in space and time dimensions using spatio-temporal templates. A simple but effective motion descriptor based on the motion-compensated frame difference is designed for template representation, which is resistant to the deformation of posture and cluttered and moving background. A multi-step filtering scheme is adopted to speed up the target candidates localization and matching to the templates. For the template sequence to video registration, we present an extended continuous dynamic programming technique which can compute the matching scores for multiple trajectories simultaneously. Extensive experimental results on different videos have demonstrated the effectiveness of the proposed method.

Keywords: Action Detection, Template Matching, Action Retrieval.

1 Introduction

Automatic detection and recognition of human action in video sequence is an extensively studied topic in computer vision and many related approaches and systems have been developed [1-4]. However, most previous work focused on constrained environments, such as video with static background [1, 2] or video with moving but relative clean background [3, 4]. When we turn to general video, some practical challenges will be encountered. One of the most difficult issues is that for general video the accurate and reliable tracking and segmentation of human body shape is usually hard to achieve. Thus it is challenging to obtain a consistent representation of human action, since that even the same action will produce different space-time intensity patterns when performed by different people wearing different clothes and in different backgrounds.

In this paper, we investigate the method for finding specific human action in videos captured with uncontrolled settings, which is important for many potential applications such as visual surveillance, content based video retrieval and sports motion analysis [5, 6]. Human action here is represented as a sequence of body

* Corresponding author.

posture, and we treat action detection and localization as matching the spatio-temporal template to the test video to detect the action's occurrence at some specific time and space location. To do this, we propose an effective feature descriptor to represent the action templates and an efficient template sequence matching method.

For human detection and action recognition, motion based scheme is mostly used because motion is not sensitive to appearance change. Efros et al. [3] developed a generic approach to recognize actions in "medium field" sports video using a motion descriptor deduced from noisy optical flow measurements. Similarly, Zhu et al. [4] proposed using slice based optical flow histograms to classify player basic action in tennis video sequence. Though promising results have been achieved, their methods need automatic tracking to get the stabilized figure-centric sequence which is impractical for general video. Recently, spatio-temporal SIFT features or SIFT trajectory [5, 7, 8] are proposed as descriptor for action recognition. Such methods do not require segmentation; however, they are not robust enough to cluttered background because these descriptors can be affected by salient background keypoints motion. Another kind of representative motion matching scheme without explicit tracking and segmentation was presented in [9], where action was localized by exhaustively test motion-consistency of small space-time patches (ST-patch) between a query video and test video. But, their method required the backgrounds of the query and test videos are both static or have same moving patterns. The approach proposed by Jiang et al. [10] can overcome this limitation, while it used edge feature to represent template which made it not robust to general background.

We propose a novel method to detect and localize human action in video. A new motion descriptor, the histogram of grid of frame-difference (*HGFD*) in human body region is proposed to compactly represent the noisy spatial motion patterns of human action. During detection and localization, we first adopt a multi-step filtering scheme like [10] to speed up search, by gradually filtering out non-object windows using motion magnitude and template clusters in each frame. Then the correspondence of object candidates in neighboring frames is constructed by tracking. Finally, an extended continuous dynamic programming (DP) method is developed to match multiple trajectories against the template sequence simultaneously and efficiently. Compared with the approaches in [9, 10], our method is robust to cluttered background and is computationally efficient. Meanwhile, it does not require the backgrounds of query and test video undergoing the same motion patterns.

2 The Proposed Method

In this section, we first introduce our motion descriptor and then present the three steps of the proposed method: target candidates localization, target candidates tracking and the template sequence matching to video.

2.1 Motion Descriptor Computation

In our approach, we utilize frame-difference (FD) to derive motion feature as it is easy to compute and independent of appearance change. Since the articulated human action is largely indicated by the relative movements of limbs in the different human regions, the frame-difference can be used to represent the spatial motion pattern of human body parts, rather than accurate segmentation of body. The histogram of grid of frame-difference (*HGFD*) in human region is used as the motion descriptor to represent templates. This descriptor captures the intrinsic features of action and is robust to moving and cluttered background since most of the background inside the human region has been eliminated in the frame difference.

We compute the difference between two successive frames and threshold it to get binary images $I_b(i, j, t)$ as follows.

$$I_b(i, j, t) = \begin{cases} 1 & | I(i, j, t) - I(i, j, t - 1) | > Th \\ 0 & otherwise \end{cases} \quad (1)$$

Then each template is divided into grids of size $m \times n$ in a sliding fashion along vertical and horizontal directions with step of $m/2$ and $n/2$ respectively. In our experiments we set $m=M/3$, $n=N/3$ and $Th=15$, where $M \times N$ is the size of template image. Thus 25 grids are obtained for each template and a histogram of 25 bins is built as the *HGFD*. The value of bin k is computed as follows.

$$bin(k) = \frac{1}{C} \sum_{(i,j) \in Grid(k)} I_b(i, j) \quad (2)$$

where C is the sum of the foreground pixels of the template. Fig. 1 shows two images of template and their respective representations.

To localize the target we need search the entire image, which means the computation cost of *HGFD* is crucial for the performance of the proposed method. To overcome this problem, we adopt the integral image introduced by Viola et al. [11] as the intermediate representation of FD, which allows the computation of the sum of the pixels within each grid very rapidly.

The similarity between two *HGFD* p, q is defined as:

$$similarity(p, q) = \sum_{u=1}^{25} \sqrt{p_u q_u} \quad (3)$$

For video recorded with moving camera, the global motion estimation and compensation for successive frames [6] are first performed before the computation of FD.

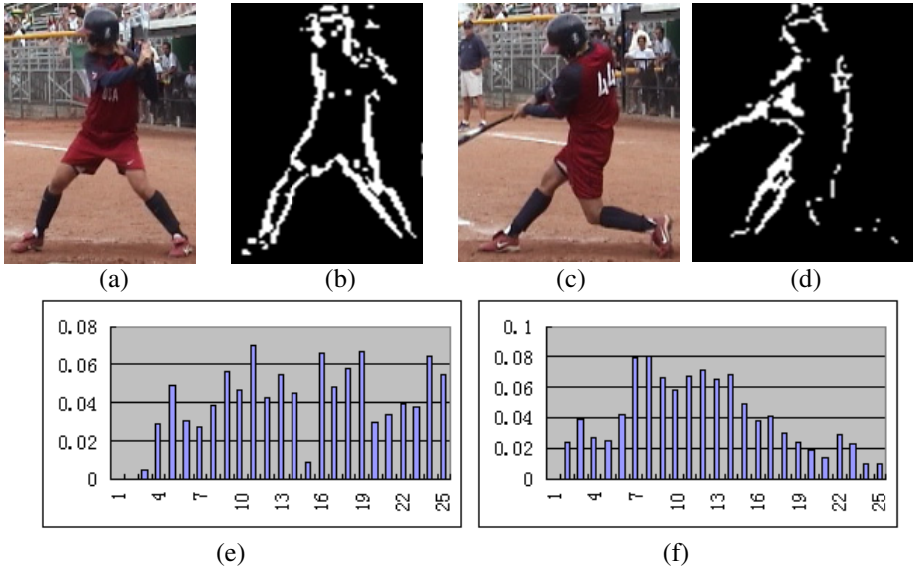


Fig. 1. Two templates (a, c), their FD (b, d) and HGFDs (e, f). For visual illumination the FD have been intensity enhanced. It can be seen that the HGFD can effectively distinguish two action templates.

2.2 Target Candidates Coarse Localization Using Motion Magnitude and Template Cluster Matching

In the first step of searching, we use the motion magnitude inside the test window and template cluster matching to coarsely localize potential targets and filter out most false targets.

Filtering Using Motion Magnitude. We measure the motion magnitude inside a test window as the sum of foreground pixels, which should be larger than a threshold Th_m . Furthermore, a test window is divided into 4 sub-window and the motion magnitude of each sub-window need be larger than Th_i ($i=1\sim 4$). Here the value for Th_m and Th_i are determined from training templates and are defined as the half of the minimal motion magnitude of template and the respective sub-template. By motion filtering, most of the test windows with no or little motion which are unlikely to contain the target are rapidly removed, reducing the cost of subsequent template matching.

Filtering Using Template Cluster Matching. Even filtered with the motion magnitude, there are still so many potential target windows in each image that it makes matching each template to them is infeasible in real applications. We propose matching using the template clusters to further eliminate the unlike candidates.

The templates in a training sequence whose motion magnitude below a value are discarded and the rest ones are clustered with k -mean algorithm. For a test window passing the motion magnitude filtering, it is further matched against the template clusters, and if the highest similarity is below a given threshold it will be deemed as clutter and removed.

For two candidate windows, if they are overlapped and the overlapped area exceeds 60% of the maximal window size, only the target with higher score is kept.

Fig.2 shows two test images and localizing results.

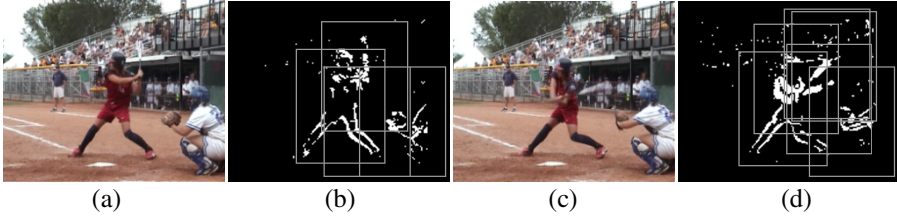


Fig. 2. Two test images (a, c) and the coarse localization results (b, d). White boxes in (b, d) are the target candidates after the two-step filtering.

2.3 Target Candidates Tracking in Neighboring Frames

After detecting the candidate targets in each frame, we need establish the correspondence between them in neighboring frames because we are matching template sequence to video. However, due to the complex background and articulated body shape, the traditional one to one color-based tracking [12] is unreliable which may miss the true target in the tracked trajectory. In our method, we keep multiple correspondences for each candidate to prevent the loss of right path based on the overlapping degree between candidate windows. If the overlapping rate of a candidate target A in current frame with a target B in previous frame exceeds 0.7, B is added to the precursor set of A. If A has no precursor in the immediate previous frame, we will check up to W previous frames until a precursor is found.

2.4 Template Sequence Matching and Action Localization

As stated in section 2.3, each target may have multiple precursors and multiple successors; hence the naïve scheme of matching the template sequence to the tracked trajectories is impractical since the total number of trajectories will grow exponentially as the template sequence length increases. In this paper, we present an extended continuous dynamic programming technique to compute the matching scores of these multiple trajectories simultaneously and efficiently.

Suppose the length of template sequence is K . We first compute the local matching distance between each target candidate and each template. For a target, O_t , in frame t , the distance to template k is defined as

$$d(o_t, k) = 1 - \text{similarity}(o_t, k) \quad (4)$$

Then the cumulative distance denoted as $S(o_t, k)$ for the best matching between the template sequence (with its end point at (o_t, k)) and the test video is computed with the extended CDP as follows.

Initial condition:

$$\left\{ \begin{array}{l} S(o_t, 0) = 2d(o_t, 0) \\ d(o_t, -1) = d(o_t, -2) = \infty \\ S(o_{t-1}, k) = S(o_t, -1) = S(o_t, -2) = \infty \end{array} \right. \quad (5)$$

Recurrence formula:

$$S(o_t, k) = \min \left\{ \begin{array}{l} \min_{o_{t-1} \in \text{Pr } \text{ior}(o_t)} (S(o_{t-1}, k)) + d(o_t, k) \\ \min_{o_{t-1} \in \text{Pr } \text{ior}(o_t)} (S(o_{t-1}, k - 1) + 2 * d(o_{t-1}, k - 1)) + d(o_t, k) \\ \min_{o_{t-1} \in \text{Pr } \text{ior}(o_t)} (S(o_{t-1}, k - 2) + 2 * d(o_{t-1}, k - 2)) + 3 * d(o_t, k) \end{array} \right. \quad (6)$$

where $\text{Prior}(o_t)$ is the precursor set of o_t . In the traditional CDP [13], only one precursor is considered so it could match only one trajectory at one time. In the extended CDP, the number of precursor is not limited and the computation complex is linear to the number of precursor. The recurrence formula in (6) allows for matching a target trajectory with a length of 1/2 to 2 times of the length of template sequence.

For a target candidate o_t , if its cumulative distance, $S(o_t, K-1)$, to the template $K-1$ is below the preset threshold then the query action is declared to have been found at frame t and at position o_t . The *viterbi* algorithm can be used to localize all the target positions in the previous frames.

3 Experimental Results

To test the effectiveness of the proposed method, four experiments with different challenges are conducted in this section.

3.1 Experiment for Softball Video

In this experiment, we test our method for detection and localization of softball batting action with complex background. The template sequence is a video clip of the player batting the ball with 10 frames. The bounding box of the player in each frame is labeled manually (see the red box in Fig.3).

The test video is another player’s same action with 102 frames. The time and space location at O_t with minimal cumulative distance $S(o_t, K-1)$ is deemed as the localization result. In the test video, the background is moving and very cluttered. Fig.3 gives some template images and localization results (indicated with blue box). Our method localizes the desired action accurately and efficiently. The total running time is 103 seconds on a P4 2.4Ghz PC with 512M of RAM. Here the resolution of

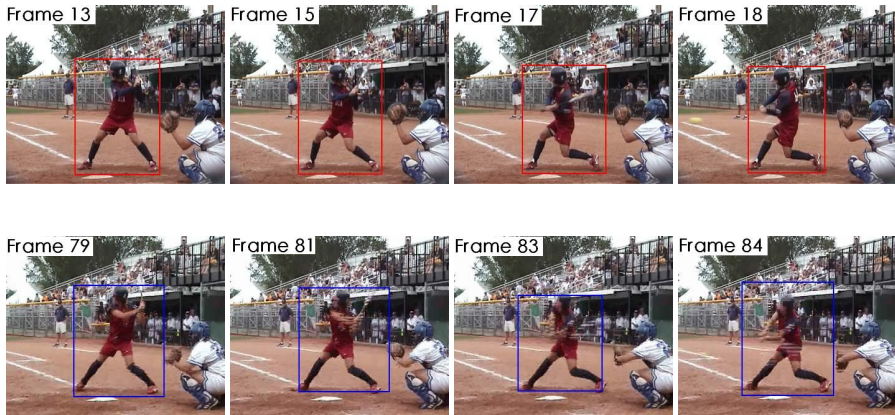


Fig. 3. Localization results for batting action. Top row: some template images. Bottom row: some action localization results.

test video is 352×288 and the average resolution of template image is 135×180 . In contrast, we extrapolated the time reported in [9] to this case, it would take about 4.5 hours.

3.2 Experiment for Walking Video

In this experiment we test our method to localize walking action. The query video is recorded with a static camera while the test video is captured with a moving camera. For this case, the method presented in [9] would fail since the background motion patterns of the query and test are different. Also, the background of the test video is much cluttered. As shown in Fig. 4, our method localizes the action and the walking pace is accurately aligned.

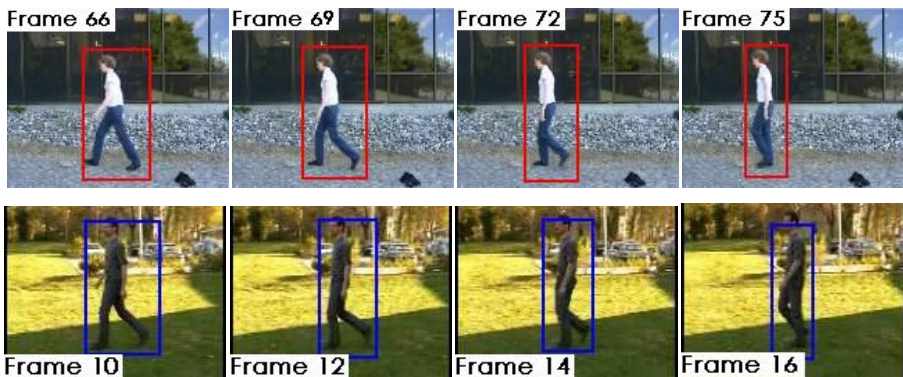


Fig. 4. Localization results for walking action. Top row: some template images. Bottom row: some action localization results.

3.3 Experiment for Diving Video

In this case, the proposed method is used to find human action with large deformation, to say, takeoff action of the diver. A women diver's platform diving is used as template and the same action of two other divers are used as test videos. As shown in Fig.5, though the diver's body changes greatly, our method can localize the action accurately.

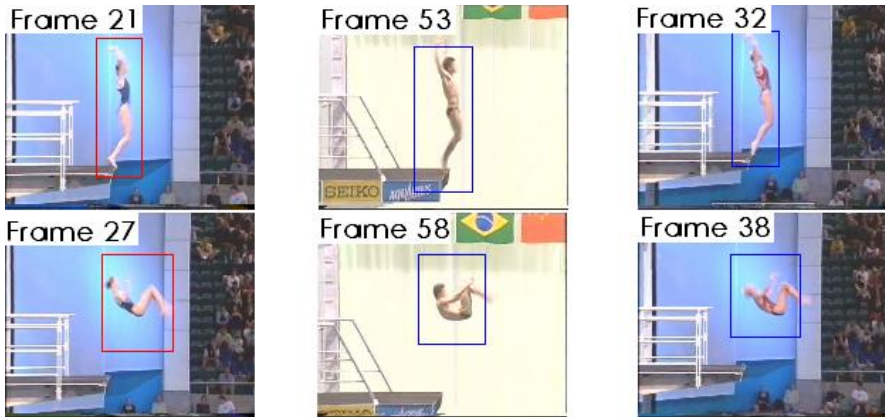


Fig. 5. Localization results for diving action. First column: template images. Second and third column: localization results.

3.4 Experiment for Baseball Video

Finally, we conduct experiment to search for a specific action, to say, throwing action in a baseball video. For the 1300 frames test video, the throwing action appears 3 times. There is large deformation and strong background clutter. Our method finds all the 3 occurrences of the action at the top of the shortlist (see Fig. 6).

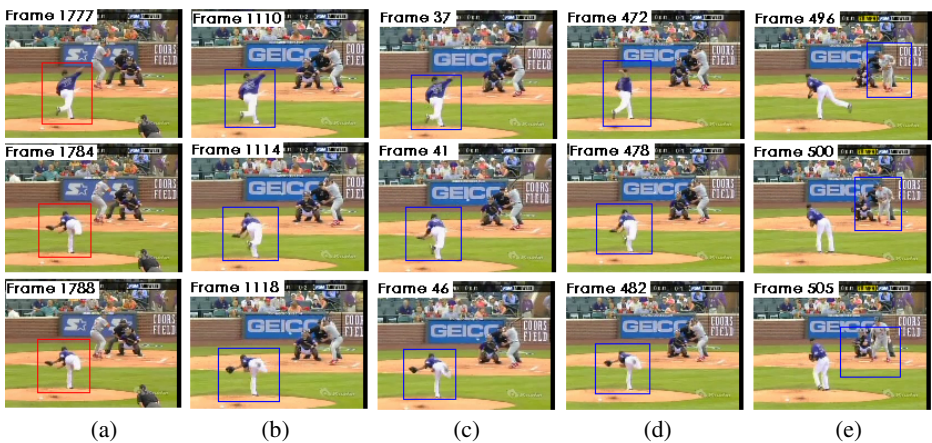


Fig. 6. Searching throwing action. (a) Templates; (b-e) Top 4 matches of the shortlist. The matching cost for (b-e) is (b): 0.158, (c): 0.178, (d): 0.197, (e): 0.464.

4 Conclusions

We have presented a robust and efficient method to accurately localize specific human action in video using spatio-temporal templates. The contributions are two-fold. The first one is the motion descriptor, *HGFD*. It reduces the influence of background pixels by the motion-compensated frame differencing, thus is insensitive to moving or cluttered background. The second one is the extended *CDP*, which allows for fast matching of template sequence to multiple trajectories to localize the action in long video. In the current implementation, we only consider actions with approximate same space scale. In the future, we will design system to handle scale-varying actions.

Acknowledgments. This work was supported by National Natural Science Funds of China (61202133, 61272214).

References

1. Bobick, A., Davis, J.: The Representation and Recognition of Action Using Temporal Templates. *IEEE Trans. on PAMI* 23(3), 257–267 (2001)
2. Yamato, J., et al.: Recognizing Human Action in Time-Sequential Images using Hidden Markov Model. In: *CVPR* (1992)
3. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: *ICCV* (2003)
4. Zhu, G.-Y., Xu, C.S., Gao, W., Huang, Q.: Action Recognition in Broadcast Tennis Video Using Optical Flow and Support Vector Machine. In: Huang, T.S., Sebe, N., Lew, M., Pavlović, V., Kölsch, M., Galata, A., Kisačanin, B. (eds.) *HCI/ECCV 2006*. LNCS, vol. 3979, pp. 89–98. Springer, Heidelberg (2006)
5. Song, Y., Zheng, Y.-T., Tang, S., Zhou, X., Zhang, Y., Lin, S., Chua, T.-S.: Localized Multiple Kernel Learning for Realistic Human Action Recognition in Videos. *IEEE Trans. Circuits Syst. Video Techn.* 21(9), 1193–1202 (2011)
6. Li, H., Tang, J., Wu, S., Zhang, Y., Lin, S.: Automatic Detection and Analysis of Player Action in Moving Background Sports Video Sequences. *IEEE Trans. Circuits Syst. Video Techn.* 20(3), 351–364 (2010)
7. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. *ACM Multimedia* (2007)
8. Sun, J., Wu, X., Yan, S., Cheong, L.F., Chua, T.-S., Li, J.: Hierarchical spatio-temporal context modeling for action recognition. In: *CVPR* (2009)
9. Shechtman, E., Irani, M.: Space-Time Behavior Based Correlation. In: *CVPR* (2005)
10. Jiang, H., Li, Z.N., Drew, M.S.: Detecting Human Action in Active Video. In: *ICME* (2006)
11. Viola, P., Jones, M.: Rapid Object Detection using a Boosted Cascade of Simple Features. In: *CVPR* (2001)
12. Yilmaz, A., Javed, O., Shah, M.: *ACM Computing Surveys* 38(4) (2006)
13. Zhang, H., Guo, Y.: Facial Expression Recognition Using Continuous Dynamic Programming. In: *ICCV Workshop on RATFGRT* (2001)

A Sparse Coding Based Transfer Learning Framework for Pedestrian Detection

Feidie Liang, Sheng Tang, Yu Wang, Qi Han, and Jintao Li

Advanced Computing Research Laboratory, Beijing Key Laboratory
of Mobile Computing and Pervasive Device, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing, China
{liangfeidie, ts, wangyu, hanqi, jtli}@ict.ac.cn

Abstract. Pedestrian detection is a fundamental problem in video surveillance and has achieved great progress in recent years. However, training a generic detector performing well in a great variety of scenes has been approved to be very difficult. On the other hand, exhausting manual labeling effort for each specific scene to achieve high accuracy of detection is not acceptable especially for video surveillance applications. In order to alleviate the manual labeling effort without sacrificing accuracy of detection, we propose a transfer learning framework to automatically train a scene-specific pedestrian detector starting from a pre-trained generic detector. In our framework, sparse coding is proposed to calculate similarities between source samples and a small set of selected target samples by using the former as dictionary. The similarities are later used to calculate weights of source samples. The weights of initially detected target samples are calculated in a similar way but using the selected target dataset as dictionary. By using these weighted samples during re-training process, our framework can efficiently get a scene-specific pedestrian detector. Our experiments on VIRAT dataset show that our trained scene-specific pedestrian detector performs well and it is comparable with the detector trained on a large number of training samples manually labeled from the target scene.

Keywords: pedestrian detection, transfer learning, sparse coding.

1 Introduction

Pedestrian detection is of great importance in video surveillance. It is a key procedure for tracking, action recognition, personal identification and unusual events detection. Significant progress has been made on pedestrian detection in the past decade [1, 2], especially the appearance-based approaches [3, 4] based on large-scale training sets became more and more popular and have achieved great success. However, it is still difficult to train a generic appearance-based pedestrian detector which works robustly in a great variety of scenes. For example, it was shown that the detection rate of the popular HOG pedestrian detector trained on the INRIA data set dropped significantly when being tested on the Caltech benchmark data set [5]. To reach high accuracy in all kinds of scenes, generic pedestrian detector not only requires a huge training set to

cover a very large variety of viewpoints, resolutions, lighting conditions, illuminations and backgrounds across different scenes which needs impractically amounts of manual labeling efforts, but also a very complex model to handle so many variations which are very hard to be found. Therefore, training and tuning a well-performed pedestrian detector for one or several particular scenes of an application is more practical. However, the required manual labeling effort for each particular scene can be time-consuming, tedious and may be not practical for some applications. In order to alleviate the manual labeling task, some solutions have been proposed in recent years. For example, [6] adopted co-training framework, [7] trained multiple classifiers based on grid structure, [8, 9] used online learning framework. However, these existing approaches are based on ad hoc rules and their trained detectors may have risk of drift during the training process [10].

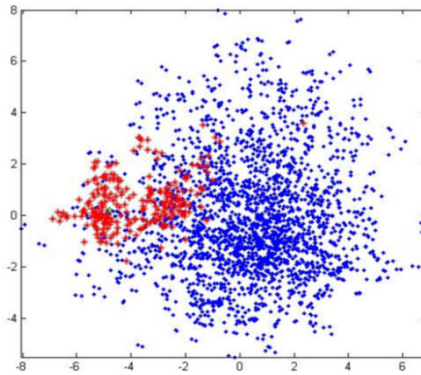


Fig. 1. Distribution of source positive samples from INRIA (blue dots) and some typical manually labeled target positive samples from VIRAT (red stars)

To tackle this problem, another emerging direction resorts to transfer learning framework to train a scene-specific pedestrian detector starting with a generic pedestrian detector pre-trained by a large set of source samples. The motivation mainly comes from the observation that although the distribution of source samples does not exactly match that of samples in a target scene, there are still some similarities between source samples and target samples, and the similarities are in varying degrees for different source samples (Fig. 1). If the source samples are weighted with different values according to these similarities in re-training process rather than treating all of them equally, the trained detector can perform better in the target scene. The challenges of this method and our contributions can be summarized as below:

1. **Weighting source samples according to visual similarities** between source samples and target samples. Distance of samples is usually taken to reflect similarities and shorter distance of two samples means that they are more similar. KNN (K-nearest-neighbor) is a popular method to calculate such distance. Based on KNN, many researches [10, 11] got good improvement in detection accuracy. However, KNN has its own inherent drawbacks. For example, a fixed global

parameter (k) is used in KNN method, which not only needs to be carefully chosen case by case to get best result, but also can be dramatically influenced by data noise and cannot handle situations where an adaptive neighborhood is required. What's more, it has been approved that sparse coding is more reliable to be used to calculate the distance than KNN in many applications such as face recognition and image classification [12, 13] as the sparse decomposition of a sample reflects its true neighborhood structure and provides a similarity measure between the sample and its neighbors. Therefore in this paper, we choose sparse coding to calculate the distances between source samples and target samples and use these distances to calculate the weight for the source samples.

2. **Selecting a small part of target samples** totally without manual labeling effort. As mentioned above, a set of selected target samples need to be used to re-weight source samples during re-training process. It is usually manually labeled from target scene in a majority of previous related works [14-17]. While some emerging approaches [18] showed that it can be automatically generated by applying some filters on the detected results of the generic detector in target scene. In our paper, context information such as motion, size and appearance are adopted as such filters in our paper.
3. **Adding weighted target samples** to the training dataset with less risk of drift. The inclusion of a big set of target samples for re-training has been approved to be essential [10]. However, it also brings risk of drift during the iterative re-training process because the automatically generated target samples are not hundred percent accurate. Label-propagation method based on KNN was adopted by [10] to solve this problem. In this paper, we propose to use sparse coding to revise the confidence value of target samples and take the revised confidence value as weight when adding these samples into the re-training dataset. Through weighting each target sample, the negative influence of adding target samples can be suppressed to a great extent.

Combining all the previous methods to solve the above challenges of transfer learning framework, we conduct a pedestrian detection prototype. The experimental results on a public video surveillance dataset show that our methods significantly improve the performance compared with the generic pedestrian detector and other re-weighting method. Moreover, the result of our approach is only slightly lower than that of manual labeling method.

2 Related Work

Early related works mainly depend on the fact that there are a small set of manually labeled samples (target samples) from the target specific scene. With this assumption, a common way to get high detection accuracy in a specific scene is training a new detector by mixing a large set of original training samples (source samples) and the target samples. Wu and Dietterich [11] demonstrated that the source samples can used to improve accuracy by taking them as auxiliary data although they are drawn from a different distribution than the target scene. Based on Adaboost, Dai et al. [14]

proposed a learning framework called TrAdaBoost to improve accuracy by decreasing the weights of wrongly predicted samples during retraining process if they belong to source dataset. The variation of view point is handled in Pang et al. [15] by a feature-representation transfer approach to adapt weights of multiple weak classifiers.

However, manual labeling work is unacceptable for some applications. Therefore, many researches turn to automatically label a small set of samples by making use of the detection results in the specific scene of original “generic” detector and different kinds of other information. Background subtraction results are used to help labeling target samples in Nair et al. [8] in its on-line learning framework. But the background subtraction results are very sensitive to lighting variations and scene clusters. Rosenberg et al. [19] selected target samples according to a training data selection metric that is defined independently of the detector rather than the selection metric that is based on the detection confidence generated by the detector. However, this method relies on the score of the selection metric which is insufficiently reliable. Wang et al. [18] integrated multiple cues of motions, path models, locations, sizes and appearance to select target samples. Its improvement mainly comes from path models, so it can only be used in scenes with paths. Some approaches are also carefully designed to tolerant the impact of labeling noise. For example, Wang et al. [10] proposed a transfer learning framework using KNN and label-propagation as tools to re-weight source samples and target samples. However, it need a complex process to converge the label-propagation [20] and the graphs constructed by KNN often fail to capture piecewise linear relationships between data samples in the same class.

3 Our Method

In this paper, we propose an instance-transfer framework based on sparse coding for pedestrian detection. It starts with a generic HOG+SVM pedestrian detector pre-trained on a generic dataset. By applying this detector on an unlabeled video sequence captured from target scene, an initial target dataset is obtained by selecting those samples with positive score. After that, a small part of samples are selected from the initial target dataset. These samples are then used to calculate the weights for the source samples and the target samples. A new detector is then trained with both weighted source samples in the generic dataset and target samples in the initial target dataset. Formally, the new detector is trained on SVM with an objective function showed as Eq. (1), where γ_i and δ_i represent the weights of source samples and target samples, which have the most importance to the performance of the new trained pedestrian detector. In this equation, w and b are weights and bias to be learned with n^s source samples and n^t target samples, which are labeled as y_i^s and y_j^t and whose HOG features are x_i^s and x_j^t respectively. C is a pre-set penalty parameter. ξ_i^s and ξ_j^t are slack variables of source samples and target samples.

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2} \|w\|^2 + C \sum_{i=1}^{n^s} (\gamma_i \xi_i^s)^2 + C \sum_{j=1}^{n^t} (\delta_j \xi_j^t)^2 \\ \text{s.t.} \quad & y_i^s (w^T x_i^s + b) \geq 1 - \xi_i^s, i = 1, \dots, n^s \end{aligned} \quad (1)$$

$$\begin{aligned} y_j^t (w^T x_j^t + b) &\geq 1 - \xi_j^t, j = 1, \dots, n^t \\ \xi_i^s &\geq 0, \xi_j^t \geq 0, \quad i = 1, \dots, n^s, j = 1, \dots, n^t \end{aligned}$$

There are many false positive samples in the initial target dataset since the generic detector does not perform well in target scene. Therefore, multiple cues such as motion, size and appearance are used in our method to filter out those false positive samples and get a small part of target samples. Based on sparse coding, these selected samples are used to calculate weights of both source samples and target samples. We will introduce how to calculate the weights in detail in the following sub-sections.

3.1 Selecting a Small Part of Target Samples of Pedestrians

When using the generic detector in target video sequence, each detection window is described with HOG feature x_j and associate with a confidence value f_j which is the output of the linear SVM classifier with the form,

$$f_j = w^T \cdot x_j + b \quad (2)$$

where w and b are the weights and bias learned by SVM with the source samples. We choose those samples with positive confidence value as the initial target samples. To further filtering out the false positive samples, three cues based on motion, size and appearance are used as the filters.

Filtering by Motion. In video surveillance applications, a detection window on a pedestrian often contains more moving pixels than that on the background. Based on this observation, most false positive samples on the background can be filtered out by a parameter defined as the ratio of moving pixels and total pixels in the detection windows. Moving pixels can be got by using background subtraction models like Vibe model [21]. According to our experiments, most false positive samples on the background can be filtered out by setting the threshold value as 0.15.

Filtering by Size. Sizes of pedestrians in a scene-specific video usually distribute in a small range, so the detected positive samples with sizes out of this range can be filtered out. As the target video is not labeled, we get the size range of normal pedestrians based on the mean size of detected positive windows. For example, we can choose the size range to cover 80% detected positive windows.

Filtering by Appearance. The generic detector is based on HOG features representing appearance, so the confidence value itself can be used as a filter. Usually detection windows on pedestrians have high confidence value. Therefore, the confidence value can be used to choose true positive target samples.

3.2 Weighting Source Samples

As shown in Fig.1, although the distribution of the source dataset usually does not match that of the samples from the target scene, some source samples better match the target dataset than others. Therefore if the source samples are assigned with different weights for re-training a scene-specific detector, the trained detector can perform better in the target scene.

Sparse coding is adopted to calculate these weights based on the selected target samples and using source dataset as bases. According to sparse coding theory, each target sample can be linearly combined by all source samples with different coefficients and most of them are zeros. After calculating this for all selected target samples, we can get an array of coefficients for each source samples, which can be seemed as matching degree between each source sample and the target dataset to some degree. Since the selected target samples are not manually labeled, some samples may be not labeled correctly. Therefore, besides this array of coefficients, the confidence value of each target samples is also included when computing the final weights. The detailed algorithm to calculate the weights is described as follows.

1. Inputs:

Source dataset $X^s = [x_1^s, x_2^s, \dots, x_{n^s}^s] \in \mathbb{R}^{d \times n^s}$

Selected target dataset $X^t = [x_1^t, x_2^t, \dots, x_{n^t}^t] \in \mathbb{R}^{d \times n^t}$

(where d is the dimension of feature; n^s and n^t is number of samples in X^s and X^t .)

2. Outputs:

Calculated weights $[\gamma_1, \gamma_2, \dots, \gamma_{n^s}]$

Sparse coding: For each sample x_j^t , solve the l^1 norm minimization problem

$$\min_{\alpha_j} \|x_j^t - D\alpha_j\|_2^2 + \lambda \|\alpha_j\|_1 \quad s. t. \alpha_j \geq 0 \quad (3)$$

where D is a dictionary of sparse representation and equal to X^s ; $\alpha_j \in \mathbb{R}^{n^s}$ and $\alpha_j \geq 0$ means all the elements of α_j are non-negative.

Normalization: Normalize each i -th element α_{ij} of α_j by

$$\alpha_{ij} = \frac{\alpha_{ij} \times \|\alpha\|_0}{\sum \alpha_{ij}}, \quad \text{for } i = 1, \dots, n^s \text{ and } j = 1, \dots, n^t \quad (4)$$

where $\|\alpha\|_0$ means the number of non-zeros for all α .

Weighting: For each source sample x_i^s , calculate the weight γ_i by

$$\gamma_i = \sum_{j=1}^{n^t} \alpha_{ij} f_j \quad (5)$$

where f_j is confidence value of x_j^t calculated by Eq. (2).

3.3 Weighting Target Samples

Detected target samples also need to be used to train the new scene-specific detector. However, they can't be used directly as they may be wrongly labeled. To tolerate the inaccuracy of target samples, we also weight them according to the similarities between them and the datasets of selected positive and false positive samples. Sparse coding method is also used to calculate the similarities by using the datasets of selected samples as bases. Our basic idea is setting larger (lower) weight if a target sample is more similar to the dataset of selected positive (negative) target samples, which is implemented by the following algorithm.

1. Inputs:

Target dataset $X^t = [x_1^t, x_2^t, \dots, x_{n^t}^t] \in \mathbb{R}^{d \times n^t}$

Selected positive dataset $X^{hp} = [x_1^{hp}, x_2^{hp}, \dots, x_{n^{hp}}^{hp}] \in \mathbb{R}^{d \times n^{hp}}$

Selected negative dataset $X^{hn} = [x_1^{hn}, x_2^{hn}, \dots, x_{n^{hn}}^{hn}] \in \mathbb{R}^{d \times n^{hn}}$

(Where d is the dimension of feature; n^t , n^{hp} and n^{hn} are number of samples in X^t , X^{hp} and X^{hn} respectively.)

2. Outputs:

Calculated weights $[\delta_1, \delta_2, \dots, \delta_{n^t}]$

Sparse coding: For each sample x_j^t , solve the l^1 norm minimization problem

$$\min_{\beta_j} \|x_j^t - D\beta_j\|_2^2 + \lambda \|\beta_j\|_1 \quad s.t. \beta_j \geq 0 \quad (6)$$

where matrix $D = [X^{hp}, X^{hn}, \mathbf{1}] \in \mathbb{R}^{d \times (n^{hp} + n^{hn} + d)}$, $\beta_j \in \mathbb{R}^{n^{hp} + n^{hn} + d}$ and $\beta_j \geq 0$ means all the elements of β_j are non-negative. X^{hp} and X^{hn} are subsets of X^t , so remove it from D and set the corresponding β_{ij} as 0 if $x_j \in D$.

Normalization: The normalization for β_{ij} is the same as Eq.(4).

Weighting: For each target sample x_j^t , calculate its weight δ_i by

$$\delta_j = f_j (\sum_{i=1}^{n^{hp}} \beta_{ij} f_i^{hp} - \sum_{k=n^{hp}+1}^{n^{hp}+n^{hn}} \beta_{kj} f_k^{hn}) \quad (7)$$

where f_j is the confidence value of x_j^t ; f_i^{hp} , f_k^{hn} is the confidence value of x_i^{hp} and x_k^{hn} respectively.

Some source and target samples with different weights are shown in Fig. 2. We can see that the results meet our expectation. These source samples with larger weights are more similar to samples of the target scene, while these source samples with small or even zero weights are obviously outliers of target scene (Fig. 2(a)). From Fig. 2(b), we can see that these detected target samples with high weights are true pedestrians, while these detected target samples with negative weights are not pedestrians.

4 Experiments

Our experiments use two public datasets: INRIA [3] and VIRAT [22]. INRIA is a public pedestrian dataset which includes lots of positive samples and negative samples taken in different scenes. We use it as the generic dataset to train a generic HOG+SVM pedestrian detector. From VIRAT, some typical videos are selected as target training and testing videos which are recorded in a same scene by a stationary camera facing a street (Fig. 4). It is challenging to accurately detect out pedestrians in such scenes because there are both moving pedestrians and vehicles with many occlusions and varying illumination conditions. We manually labeled 2000 positive samples and 12000 negative samples from 5500 frames of target training videos, which are used to train a scene-specific detector for comparison. Note that these manually labeled samples are thoroughly not used to train our scene-specific detector.

PASCAL criterion is adopted to distinguish true positive windows from false positive ones. A detected positive window is treated as correct only if area overlap of this window and a ground truth window exceeds 50%. We use DET curve that Miss Rate versus False Positive Per Image (FPPI) as the evaluation metric. For DET curve, lower curve means better performance of detection. In the following description, we assume that FPPI is 0.1 when we talk about detection rate if FPPI is not specified.



Fig. 2. Some example samples with different weights (a) Positive source samples from INRIA dataset with large weight (first row) and zero weight (second row). (b) Positive target samples from VIRAT dataset with large weight (first row) and negative weight (second row).

4.1 Overall Performance

We compare the detection results on target testing videos of the scene-specific detector trained by our approach with two other detectors: one is the generic detector trained on the INRIA dataset (Generic), the other is a scene-specific detector trained on a large number of manually labeled samples from the target training set (Manual). The DET curves of these three detectors on VIRAT are showed in Fig. 3(a). Note all the DET curves are based on the results after applying the motion cue on the detected positives.

From Fig.3(a), we can see that the scene-specific detector obtained by our approach significantly outperforms the generic detector, it improves the total detection performance¹ from 71% to 48%. The detection rate of our detector is about 52%, while the detection rate of the generic detector is only about 24%. The results verify that our scene-specific detector is much more suitable for the specific scene than the generic detector. We can also see that our scene-specific is comparable with the

¹ The value is list in the Fig, lower value means better detection.

manual scene-specific detector. The detection performance of the manual scene-specific detector is 46% and the detection rate is about 57%, which is only slightly better than that of our scene-specific detector.

As a result, from the detection results shown in Fig. 4, we can see that most pedestrians in the target testing videos can be detected out.

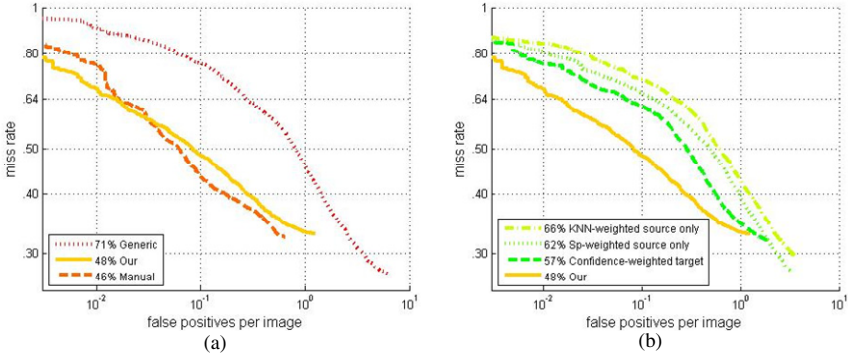


Fig. 3. DET curves on VIRAT testing videos of different detectors (a) Overall performance comparison of our approach to a generic detector and a scene-specific trained on a large number of manually labeled samples from the target scene. (b) Impacts of different strategies on detection performance.

4.2 Impacts of Different Strategies on Detection Performance

Besides our approach, there are many other strategies about how to train a scene-specific detector based on a generic one. We select three typical strategies and compare their detection performance: (1) Sp-weighted source only: only use source samples weighted by our method in section 3.2. (2) KNN-weighted source only: only use source samples weighted by KNN method introduced in [10]. (3) Confidence-weighted target: same as our approach but use confidence value of detection as weights of target samples.

We can see from Fig. 3(b) that our method based on sparse coding is more appropriate than KNN-based method to get the weights of the source samples. When only source samples are included in the re-training process, the detection performance is 62% and the detection rate is about 34% when the source samples are weighed by our method, while these numbers are 66% and 29% respectively for KNN-based method. This is mainly because sparse coding catches the piecewise linear relationship between source samples and target samples rather than the pairwise relationship caught by KNN, so it can perform better than KNN to map source samples into target scene.

Another conclusion from Fig. 3(b) is that target samples can be used to help improve the detection performance of retrained scene-specific detector and our method performs better than directly using confidence value as the weight. By adding the target samples weighted by confidence value, the detection performance can be improved from 62% to 57%, and it is further improved to 48% if the target samples are weighted by our approach. The main reason is that the detection confidence values of

the generic detector are not absolutely accurate and the correctness of re-trained detector may be influenced by these imprecise weights. For example, there are usually some false positive samples with high confidence value in the initial detection results. In our approach, the weights of the target samples are calculated based on sparse coding between the target samples and selected positive and negative samples, so the impact of such inaccuracy can be avoided to some extent.

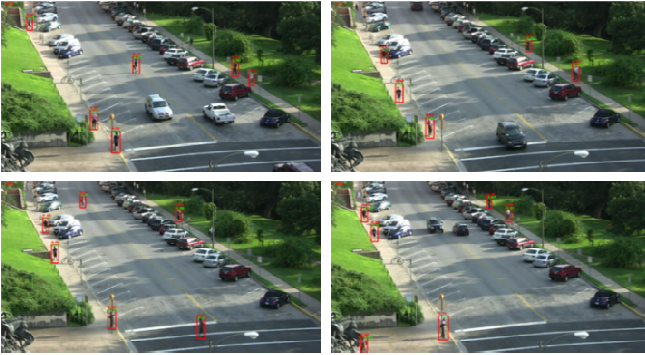


Fig. 4. Some detection results of our approach on VIRAT testing videos

5 Conclusion

In this paper, we propose a transfer learning framework based on sparse coding to train a scene-specific pedestrian detector. It is based on a generic pre-trained detector but don't need any manual labeling efforts in target scene. The main idea is to weight the samples when they are used to re-train the scene-specific detector according to the similarities between these samples and a small set of selected positive samples and negative samples. These selected samples are obtained by applying multiple context cues on the initial detection results. Sparse coding is introduced to calculate such visual similarities, which is approved to be more efficient than other methods like KNN. Our experiments on a public dataset show that the scene-specific detector trained by our approach significantly outperforms the generic pedestrian detector. It is also comparable with the pedestrian detector trained with a number of manually labeled samples from target scene. In the future, we plan to adopt parallel strategies [23] into our method to improve the efficiency.

Acknowledgement. This work was supported in part by the National Nature Science Foundation of China (61173054, 61271428); and Co-building Program of Beijing Municipal Education Commission.

References

1. Dollar, P., et al.: Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(4), 743–761 (2012)

2. Munder, S., Gavrilu, D.: An Experimental Study on Pedestrian Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI 28(11) (2006)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, pp. 886–893 (2005)
4. Felzenszwalb, P., McAllester, D., Ramanan, D.: A Discriminatively Trained, Multiscale, Deformable Part Model. In: *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR (2008)
5. Dollár, P., et al.: Pedestrian detection: A benchmark. In: *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, pp. 304–311 (2009)
6. Levin, A., Viola, P., Freund, Y.: Unsupervised improvement of visual detectors using co-training. In: *IEEE International Conference on Computer Vision* (2003)
7. Roth, P.M., et al.: Classifier grids for robust adaptive object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, pp. 2727–2734 (2009)
8. Nair, V., Clark, J.J.: An unsupervised, online learning framework for moving object detection. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2004)
9. Bo, W., Nevatia, R.: Improving Part based Object Detection by Unsupervised, Online Boosting. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2007)
10. Meng, W., Wei, L., Xiaogang, W.: Transferring a generic pedestrian detector towards specific scenes. In: *IEEE Computer Conference on Computer Vision and Patter Recognition* (2012)
11. Wu, P., Dietterich, T.G.: Improving SVM accuracy by training on auxiliary data sources. In: *Proceedings of the Twenty-First International Conference on Machine Learning*, p. 110. ACM, Banff (2004)
12. Bin, C., et al.: Learning With l1-Graph for Image Analysis. *IEEE Transactions on Image Processing* 19(4), 858–866 (2010)
13. Tang, S., et al.: Sparse Ensemble Learning for Concept Detection. *IEEE Transactions on Multimedia* 14(1), 43–54 (2012)
14. Dai, W., et al.: Boosting for transfer learning. In: *Proceedings of the 24th International Conference on Machine Learning*, pp. 193–200. ACM, Corvallis (2007)
15. Junbiao, P., et al.: Transferring Boosted Detectors Towards Viewpoint and Scene Adaptiveness. *IEEE Transactions on Image Processing* 20(5), 1388–1400 (2011)
16. Wang, M., et al.: Assistive Tagging: A Survey of Multimedia Tagging with Human-Computer Joint Exploration. *ACM Computing Surveys* 44(4) (2012)
17. Wang, M., et al.: Towards a Relevant and Diverse Search of Social Images. *IEEE Transactions on Multimedia* 12(8), 829–842 (2010)
18. Meng, W., Xiaogang, W.: Automatic adaptation of a generic pedestrian detector to a specific traffic scene. In: *IEEE Computer Conference on Computer Vision and Patter Recognition* (2011)
19. Rosenberg, C., Hebert, M., Schneiderman, H.: Semi-Supervised Self-Training of Object Detection Models. In: *IEEE Workshops on Application of Computer Vision* (2005)
20. Wright, J., et al.: Sparse Representation for Computer Vision and Pattern Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 98(6) (2010)
21. Barnich, O., Van Droogenbroeck, M.: ViBe: a universal background subtraction algorithm for video sequences. *IEEE Transactions on Image Process*, ITIP 20(6) (2011)
22. Sangmin, O., et al.: A large-scale benchmark dataset for event recognition in surveillance video. In: *IEEE Conference on Computer Vision and Pattern Recognition* (2011)
23. Zhang, Y., et al.: Efficient Parallel Framework for H.264/AVC Deblocking Filter on Many-core Platform. *IEEE Transactions on Multimedia* (2012)

Sampling of Web Images with Dictionary Coherence for Cross-Domain Concept Detection

Yongqing Sun, Kyoko Sudo, Yukinobu Taniguchi, and Masashi Morimoto

NTT Media Intelligence Laboratories,
1-1 Hikarinooka Yokosuka-shi Kanagawa, 239-0847, Japan
yongqing.sun@lab.ntt.co.jp

Abstract. Due to the existence of cross-domain incoherence resulting from the mismatch of data distributions, how to select sufficient positive training samples from scattered and diffused web resources is a challenging problem in the training of effective concept detectors. In this paper, we propose a novel sampling approach to select coherent positive samples from web images for further concept learning based on the degree of image coherence with a given concept. We propose to measure the coherence in terms of how dictionary atoms are shared since shared atoms represent common features with regard to a given concept and are robust to occlusion and corruption. Thus, two kinds of dictionaries are learned through online dictionary learning methods: one is the concept dictionary learned from key-point features of all the positive training samples while the other is the image dictionary learned from those of web images. Intuitively, the coherence degree is then calculated by the Frobenius norm of the product matrix of the two dictionaries. Experimental results show that the proposed approach can achieve constant overall improvement despite cross-domain incoherence.

Keywords: Visual concept detection, Semantic indexing, Web image mining, Sparse representation, Dictionary learning.

1 Introduction

Nowadays, the explosive growth of visual contents on the Internet presents a challenge in how to manage the ever-growing size of the multimedia collections, particularly in how to extract sufficiently accurate semantic metadata (concepts) to make them searchable [5]. Visual concept detection is essentially a classification task in which classifiers are learned with various features extracted from training samples to predict the presence of a certain concept in a video shot or keyframe (image) [14, 15]. Ranging from objects such as “boat” and “car” to scenes such as “sky” and “sea”, semantic concepts can serve as good intermediate semantic metadata for video content indexing and understanding [14]. With a large set of robust concept detectors, significant improvement can be achieved in many challenging applications, such as image/video search and summarization [15].



Fig. 1. Web Image Examples of “Airplane-flying”

In order to learn effective concept detectors, a critical step is to acquire a sufficiently large amount of training samples, especially positive training samples [5]. Fortunately, with the explosive growth of visual contents on the Internet, large amounts of training samples have become available through Web searching [2, 15]. Consequently, how to utilize these abundant web images to improve concept detection has been the subject of intensive research by a large multimedia research community, since it has offered promising ways to automatically annotate the contents at relatively low cost [2, 15]. [15] empirically studied the effect of exploiting tagged images on concept learning by analyzing tag lists. [2] proposed an automatic concept-to-query mapping method for acquiring training data from online platforms.

However, the online web images are very noisy, cover a wide range of unpredictable contents, and have quite different data distributions with any close dataset such as TREC-Vid dataset [9, 13]. As shown in Figure 1, for example, the content of web images searched from Google Image with the keyword “Airplane-flying” varies greatly. Obviously, the images in the top row of the figure are incoherent from the concept “Airplane-flying” in the TRECVID dataset. Thus these images can not facilitate the training of the concept and may even harm it. Only the images in the bottom row are consistent with the dataset and hence helpful. Therefore, how to select coherent positive training samples from diffused web images is a challenging problem for training of effective concept detectors [2, 11, 12] due to the existence of cross-domain incoherence resulting from the mismatch of data distributions.

Existing work on video concept learning using web images has mainly focused on how to leverage compact features, such as region-based features [12] or image saliency [11], to alleviate the visual differences. Since an image is greatly reduced to a very compact feature vector, the effect of these approaches is not evident. In this paper, we propose a novel sampling approach on how to exploit bundles of local key-point features to measure how coherent a web image is with a given concept, from the aspects of sparse coding and dictionary learning.

2 Sparse Coding and Dictionary Learning

Recently, modeling data or signals as sparse linear combinations of a few elements (atoms) of some redundant bases (dictionary), sparse coding or sparse representation has been widely applied to classification problems where the data on multiple subspaces relies on the notion of sparsity due to its robustness to occlusion and corruption [8].

Formally, given a dictionary $\mathbf{D} = [d_1, \dots, d_k] \in \mathbf{R}^{l \times k}$, and the i -th data instance vector $x_i \in \mathbf{R}^l$ from the observed data matrix $\mathbf{X} = [x_1, \dots, x_n] \in \mathbf{R}^{l \times n}$, where d_i is the dictionary atom, l is the data dimensionality, k is the size of the dictionary, and n is the number of data instances ($l < k \ll n$), sparse representation solves the following non-convex program to seek the sparsest solution for the coefficient vector (i.e., sparse code) $\alpha_i \in \mathbf{R}^k$:

$$\min_{\alpha_i} \|\alpha_i\|_0 \quad s.t. \quad \mathbf{D}\alpha_i = x_i \quad (1)$$

where $\|\alpha_i\|_0$ denotes the l_0 pseudo-norm of the coefficient vector $\alpha_i \in \mathbf{R}^k$, i.e., the number of non-zero elements.

Since minimizing l_0 is NP-hard, a common approximation is to replace it with the l_1 -norm according to theories from compressive sensing [3]. Taking noise into consideration, the equality constraint must be relaxed. Hence, an alternative is to solve the unconstrained problem after using the Lagrange multiplier method:

$$\min_{\alpha_i} \frac{1}{2} \|x_i - \mathbf{D}\alpha_i\|^2 + \lambda \|\alpha_i\|_1, \quad (2)$$

where λ is a regularization parameter that balances the tradeoff between reconstruction error and sparsity induced by the alternative l_1 -norm constraint. This is a convex problem called Lasso in statistics and can be efficiently solved by the LARS-Lasso algorithm [4].

Here, how to determine the dictionary \mathbf{D} is very important for sparse representation. It has been shown that dictionaries learned from data can significantly outperform off-the-shelf ones such as wavelets [8]. Given the observed data matrix \mathbf{X} , the goal is to seek an optimal \mathbf{D} so that all the data instances can be represented as a sparse linear combination of their atoms. There are many dictionary learning methods such as the method of optimal directions (MOD), the K-SVD algorithm, and the Generalized Principal Component Analysis (GPCA) [8]. All the methods using classical optimization alternate between the dictionary and sparse code, and can obtain good results, but are too slow to scale up to large data sets [8]. Recently, efficient online learning methods were proposed in [8], which can handle large scale, potentially infinite, or dynamic data sets.

3 Proposed Approach

3.1 Overview

Inspired by the observation that dictionary atoms representing common features in all categories tend to appear to be repeated almost exactly in dictionaries

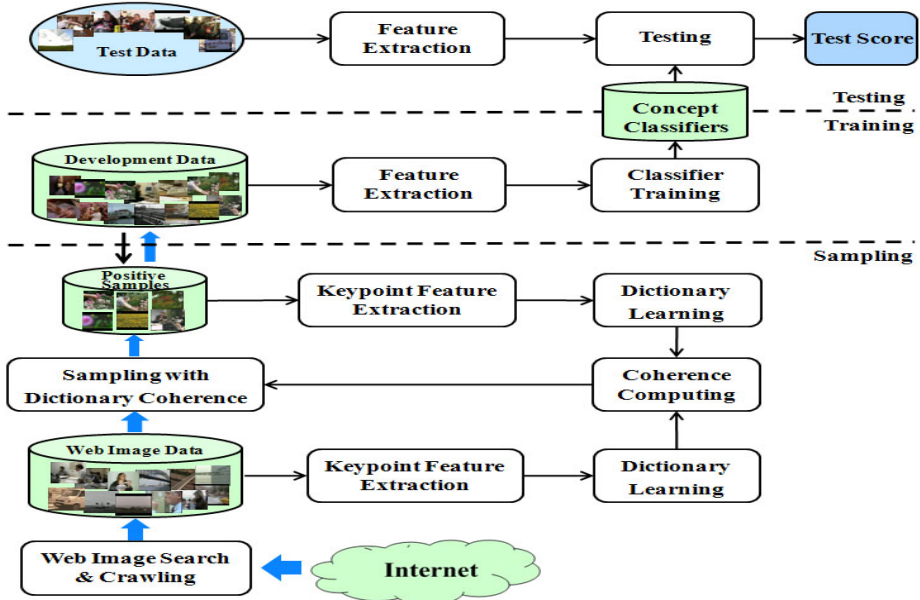


Fig. 2. Proposed Framework

corresponding to different categories, [10] promotes incoherence between the dictionary atoms to improve the speed and accuracy of sparse coding.

Motivated by this work, since the shared dictionary atoms learned from data can represent common features with regard to a given concept (represented by the set of positive training samples) and are robust to occlusion and corruption [8], we propose to use dictionary coherence in terms of how an image and a given concept share dictionary atoms to measure the degree of image coherence with the concept. That is, the more atoms they share, the higher the dictionary coherence is, which means it is more probable that the web image is coherent with the concept.

In order to compute the dictionary coherence, we learn two kinds of dictionaries through the online dictionary learning method [8]: one is the concept dictionary learned from key-point features of all the positive training samples while the other is the image dictionary learned from those of web images. Intuitively, the coherence degree is then calculated by the Frobenius norm of the product matrix of the two dictionaries since it reflects the sum of the absolute values of inner products between dictionary atoms.

On the basis of the dictionary coherence, we propose a novel adaptive sampling approach to select coherent positive samples from diffused web images for further concept learning.

3.2 Algorithm

As shown in the framework of Figure 2, for each concept, the algorithm of the proposed sampling principally consists of the following steps:

- (1) **Construction of concept set:** Select all the positive training samples from a development dataset such as TRECVID development set to represent the concept.
- (2) **Feature extraction of concept set:** Extract local key-point features, such as SIFT [7] or SURF [1], and collect each key-point feature $x_i \in \mathbf{R}^l$ of all the images in the concept set to form the data matrix $\mathbf{X}_C = [x_1, \dots, x_n] \in \mathbf{R}^{l \times n}$. Here, l is the feature dimensionality, and n is the total number of keypoints.
- (3) **Concept dictionary learning:** Adopt the efficient online dictionary learning methods [8] to learn the concept dictionary $\mathbf{D}_C \in \mathbf{R}^{l \times k}$ from the concept data matrix \mathbf{X}_C , where k is the size of the dictionary, i.e., the number of atoms. For the SIFT feature, we set $k = 192$ about 1.5 to 2.0 times of the feature size $l = 128$ [14].
- (4) **Collection of web image set:** After query construction or mapping [2] based on the concept name, search the web images and crawl the top-ranked ones.
- (5) **Feature extraction of web image:** For each image in the web image set, extract the same local key-point features as the second step, and form the image data matrix $\mathbf{X}_i \in \mathbf{R}^{l \times m}$, where m is the number of keypoints in the image.
- (6) **Image dictionary learning:** Adopt the same dictionary learning methods [8] to learn the image dictionary $\mathbf{D}_i \in \mathbf{R}^{l \times k}$ from the image data matrix \mathbf{X}_i .
- (7) **Dictionary coherence computing:** Use Equation (4) in subsection 3.4 to compute the dictionary coherence C_i between the image dictionary \mathbf{D}_i and the concept dictionary \mathbf{D}_C .
- (8) **Adaptive sampling:** Compare the dictionary coherence C_i of the current web image with the adaptive threshold in subsection 3.5 to determine whether to add the current web image to the training set.

As shown in Figure 2, after adding the selected coherent positive web samples (a manual check is advised to ensure it is positive) to the training set, we can do further concept learning for training more effective concept detectors. We will detail the key procedures in the following subsections.

3.3 Dictionary Learning

In our study, we use the efficient online learning methods [8] to learn the dictionary. Due to the advantage of non-negativity constraints in learning part-based representations [14], which is helpful for object-oriented concept learning, we impose the positivity constraints on both dictionary D and sparse code α_i in solving the optimization problem as below:

$$\min_{\mathbf{D}, \alpha_i} \sum_{i=1}^n \left(\frac{1}{2} \|x_i - \mathbf{D}\alpha_i\|^2 + \lambda \|\alpha_i\|_1 \right), \text{ s.t., } \mathbf{D} \geq 0, \alpha_i \geq 0. \quad (3)$$

while restricting the atoms to have a norm of less than one. The optimization is achieved through an iterative approach consisting of two alternative steps: the sparse coding step on a fixed \mathbf{D} and the dictionary update step on fixed α_i [8]. As mentioned above, we learn two types of dictionaries: (1) a concept dictionary \mathbf{D}_c ; (2) an image dictionary \mathbf{D}_i .

3.4 Dictionary Coherence Computing

The natural way to measure the degree of coherence C_i between the image dictionary \mathbf{D}_i and the concept dictionary \mathbf{D}_c , is to inspect the product matrix: $\mathbf{D}_i^T \mathbf{D}_c$, where the superscript T denotes the matrix transposition. This is because the element d_{ij} of the product matrix represents the inner product between a pair of the two dictionary atoms, i.e., $d_{ij} = d_i \cdot d_j$, here, $d_i \in \mathbf{D}_i$, $d_j \in \mathbf{D}_c$. Therefore, as shown in Equation (4), we compute dictionary coherence C_i through a Frobenius norm defined as the square root of the sum of the absolute squares of the matrix's elements d_{ij} :

$$C_i = \|\mathbf{D}_i^T \mathbf{D}_c\|_F = \sqrt{\sum_{i=1}^k \sum_{j=1}^k |d_{ij}|^2} \quad (4)$$

where the subscript F denotes the Frobenius norm.

3.5 Adaptive Sampling

After computing the dictionary coherence C_i between the current web image and the concept, we can easily determine whether to add the current web image to the training set by simply comparing the C_i with a pre-given threshold C_{th} . If $C_i \geq C_{th}$, meaning that the web image is coherent with the concept, then we accept it. Otherwise, we discard it.

Here, we propose an adaptive off-line method through automatic calculation of the threshold C_{th} from the distribution of the coherence degrees of all the positive train samples. According to the theory of hypothesis testing, the threshold C_{th} can be adaptively determined by:

$$C_{th} = \mu - \eta\sigma, \quad (5)$$

where μ and σ are the mean and standard deviation of all the coherence degrees C_{Pos} between each positive training sample and the concept, and η is an empirical parameter that can be determined universally. In our experiments, we set $\eta = \sqrt{3}$.

4 Experimental Results

We tested the proposed method on the TRECVID 2008 [9, 13]. TRECVID is now widely regarded as the actual standard for performance evaluation of concept based video retrieval systems [14]. The number of positive training samples

Table 1. The number of positive samples for 20 concepts in TRECVID 08

ID	Concept	#DPos	#WPos	#SPos
1001	Classroom	241	790	347
1002	Bridge	186	420	235
1003	Emergency-Vehicle	103	151	11
1004	Dog	136	795	123
1005	Kitchen	289	537	174
1006	Airplane-flying	80	395	113
1007	Two-people	4140	729	458
1008	Bus	106	902	312
1009	Driver	302	489	157
1010	Cityscape	331	879	623
1011	Harbor	217	261	76
1012	Telephone	203	557	412
1013	Street	1799	693	508
1014	Demonstration-Or-Protest	159	68	25
1015	Hand	1879	384	302
1016	Mountain	265	507	284
1017	Nighttime	490	594	229
1018	Boat-Ship	506	783	215
1019	Flower	620	948	513
1020	Singing	441	646	187

Note: The column “#DPos” denotes the number of positive training samples in the TRECVID 08 development set, “#WPos” in the initial positive web image set, “#SPos” in the final web image set after sampling.

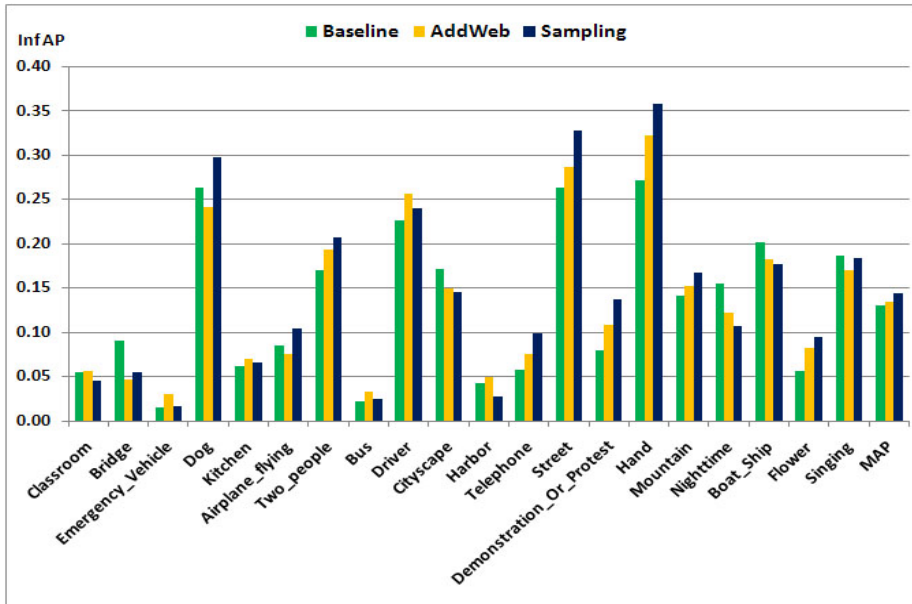


Fig. 3. Comparison results

for each concept in the TREC Vid 08 development set is shown in the column “#DPos” of Table 1 [14].

First, we used the Google API to search and download the top 1000 web images for each concept by constructing a query with the concept name. Then we annotated the images manually; the number of positive samples for each concept in the initial web image set is shown in the column “#WPos” of Table 1. Finally, we used our proposed sampling method to select the positive samples for each concept; the number of positive samples for each concept selected from the web images is shown in the column “#SPos” of Table 1. To test the effectiveness of our proposed method, we performed three runs for each concept:

- **[Baseline]**: Use only positive training samples in the TREC-Vid 08 development set (“#DPos” in Table 1).
- **[AddWeb]**: Use positive training samples of the TREC-Vid 08 development set and the initial positive web image set (“#DPos+#WPos” in Table 1).
- **[Sampling]**: Use positive training samples of the TREC-Vid 08 development set and the web image set after the proposed sampling (“#DPos+#SPos” in Table 1).

In the above runs, we used the SIFT features [7] for dictionary learning during sampling, and the well-known BoW feature [6] based on soft-weighting of SIFT, due to its widely reported effectiveness [14].

Figure 3 shows the comparison results of AP for each concept and mean AP (MAP) of the three runs. As shown, the proposed run [Sampling] achieved

the highest MAP of 0.144, which is 9.92% higher than the run [Baseline] (MAP 0.131), and 6.67% higher than the run [AddWeb] without sampling (MAP 0.135). In particular, the proposed method outperformed the others on 9 out of 20 concepts, including Airplane-flying, Dog, Telephone, Demonstration-Or-Protest, Hand, and Flower, which had been selected with sufficient visually-coherent positive samples, while little was gained with the concepts such as Harbor, Kitchen, Bridge, and Emergency-Vehicle because these concepts on the old documentary TRECVID videos may be too outdated for enough positive web samples to be obtained. On the other hand, the run [AddWeb] achieved only a 3.05% improvement in MAP compared with the run [Baseline].

Compared with the best runs in TRECVID 2008 [6], significant improvement was obtained in handling concepts with few TRECVID positive training samples. The experimental results show that the proposed approach can achieve constant overall improvement despite cross-domain incoherence.

5 Conclusion

In this paper, we proposed a novel sampling of web images for cross-domain concept detection based on coherence between an image dictionary and concept dictionary. Experimental results on the TRECVID 08 benchmark show the effectiveness and necessity of the proposed sampling method.

References

1. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. *Computer Vision and Image Understanding* 110(3), 346–359 (2008)
2. Borth, D., Ulges, A., Breuel, T.M.: Automatic concept-to-query mapping for web-based concept detector training. In: *ACM Multimedia 2011, New York, USA*, pp. 1453–1456 (2011)
3. Donoho, D.: For most large underdetermined systems of linear equations the minimal l_1 -norm solution is also the sparsest solution. *Comm. Pure and Applied Math.* 59(6), 797–826 (2006)
4. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. *Annals of Statistics* 32(2), 407–499 (2004)
5. Huiskes, M.J., Thomee, B., Lew, M.S.: New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative. In: *MIR 2010, New York, USA*, pp. 527–536 (2010)
6. Jiang, Y.-G., Yang, J., Ngo, C.-W., Hauptmann, A.G.: Representations of keypoint-based semantic concept detection: A comprehensive study. *IEEE Transactions on Multimedia* 12, 42–53 (2010)
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
8. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.* 11, 19–60 (2010)
9. Over, P., Awad, G., Rose, R.T., Fiscus, J.G., Kraaij, W., Smeaton, A.F.: Trecvid 2008 - goals, tasks, data, evaluation mechanisms and metrics. In: *TRECVID Workshop (2008)*

10. Ramirez, I., Sprechmann, P., Sapiro, G.: Classification and Clustering via Dictionary Learning with Structured Incoherence and Shared Features. In: CVPR 2010, pp. 3501–3508 (June 2010)
11. Sun, Y., Kojima, A.: A novel method for semantic video concept learning using web images. In: ACM Multimedia 2011, New York, USA, pp. 1081–1084 (2011)
12. Sun, Y., Shimada, S., Taniguchi, Y., Kojima, A.: A novel region-based approach to visual concept modeling using web images. In: ACM Multimedia 2008, New York, USA, pp. 635–638 (2008)
13. Tang, S., Li, J.-T., Li, M., Xie, C., Liu, Y.-Z., Tao, K., Xu, S.-X.: TRECVID 2008 High- Level Feature Extraction By MCG-ICT-CAS. In: Proc. TRECVID 2008 Workshop, Gaithersburg, USA (November 2008)
14. Tang, S., Zheng, Y.-T., Wang, Y., Chua, T.-S.: Sparse ensemble learning for concept detection. *IEEE Transactions on Multimedia* 14(1) (2012)
15. Zhu, S., Wang, G., Ngo, C.-W., Jiang, Y.-G.: On the sampling of web images for learning visual concept classifiers. In: CIVR 2010, New York, USA, pp. 50–57 (2010)

Weakly Principal Component Hashing with Multiple Tables

Haiyan Fu, Xiangwei Kong, Yanqing Guo, and Jiayin Lu

School of Information and Communication Engineering
Dalian University of Technology, Dalian, China, 116023
{fuhy,kongxw,guoyq}@dlut.edu.cn, hllujiayin@yahoo.com.cn

Abstract. Image hashing based Approximate Nearest Neighbor (ANN) searching has drawn much attention in large-scale image dataset application, where balance the precision and high recall rate is difficulty task. In this paper, we propose a weakly principal component hash method with multiple tables to encode binary codes. Analyzing the distribution of projected data on different principal component directions, we find that neighbors which are far in some principal component directions maybe near in the other directions. Therefore, we construct multiple-table hashing to search the missed positive samples by previous tables, which can improve the recall. For each table, we project data to different principal component directions to learn hashing functions. In order to improve the precision rate, neighborhood points in Euclidean space should also be neighborhoods in Hamming space. So we optimize the projected data using orthogonal matrix to preserve the structure of the data in the Hamming space. Experimental and compared with six hashing results on public datasets show that the proposed method is more effective and outperforms the state-of-the-art.

Keywords: Approximate Nearest Neighbor Search, Image Hashing, Weakly Principal Component Hashing, Multiple Hashing Tables.

1 Introduction

The aim of nearest neighbor search[26][14][15][16][19][21][29][28] is to find the most similar points to the query point from the image database. When database containing even billions of samples, nearest neighbor search is infeasible. Moreover, the data dimensionality is as high as hundreds or thousands, so the storage and retrieval speed will become a hardwork in scalable dataset. Therefore, in some applications it may be acceptable to retrieve an approximate nearest neighbor which doesn't guarantee to return the actual nearest neighbor in every case, but with improved speed or memory savings[17][18][20][22]. Hashing-based methods provide a feasible way for approximate nearest neighbor searching.

The main idea of image hashing is to encode high-dimensional image descriptors into compact hash codes via hash functions so as to guarantee similar images have the same or similar hash codes. Recently, a considerable amount of relative

research has been reported on image hashing[27][2][3][4][5][6][7][10][13]. Based on the number of hash table, we group these methods into two categories: single hash table methods and multiple hash tables methods.

Single hash table methods generate short compact hash codes to represent image in one table. References[5][6][7][11][24] are belong to single hash table methods. Spectral hashing(SH)[6] produces hash codes by thresholding with nonlinear functions along the principal directions of the data. The performance of SH is well for short hash codes, but may decrease substantially when most of the data variance is contained in top few principal directions[8]. In reference [7], Wang proposes a semi-supervised hashing (SSH)utilizing labeled data. Compared with simple metrics, SSH defines a pairwise label matrix based on a few pairwise labels to express the semantic similarity well. Under the SSH framework, Wang et al.[8] also proposes sequential projection learning hashing (SPLH) to correct the "errors" produced by the previous hashing function. In reference [9], Gong et al. formulates the hashing code learning problem in terms of directly minimizing the quantization error of mapping the PCA projected data to vertices of the binary hypercube. It is proved that balancing the variance of different PCA directions can achieve better results than SH and SSH. Single hash table methods retrieve all the points using a Hamming ball centered at the query, in a way that the Hamming distances between the returned points and the query are not longer than the radius of this Hamming ball. Since higher recall needs large radius of the Hamming ball, which will result in low precision[1].

In general, multiple hash tables methods generate L hash tables with K hash-codes in each table. Locality-sensitive hashing (LSH) and its extension version [2][3][4][10][12]are the representative multiple hash tables methods. Because random vectors with some specific distribution are used as the projection bases in LSH, it uses multiple hash tables to ensure that the probability of collision is much higher for points close to each other than those far away. Since LSH is data-blind and independent, it needs long codes to guarantee the collision probability. As a result, lots of hash tables are needed to get a good recall, which leads to the heavy increase of storage requirement for large scale applications[5]. Different from LSH with multiple random hash tables, Xu et al.[1]proposed complementary hashing(CH) to balance the precision and recall rate using multiple learned hash tables. CH constructed the hash functions in a complementary manner, such that the positive points of query sample lost from one hash table would be searched by the following hash tables.

In order to reach high precision and recall, in this paper, we propose a weakly principal component hashing with multiple tables. Analyzing the PCA projected data distribution, we find that the projected data in different principal component directions together can help to search more positive neighbors. Therefore, we consider these weakly principal components and learn multiple-table hashing functions, which will help to retrieve the missing samples by the previous tables and increase the number of neighbors, as well as improve the precision. For each hash table, we maximize the variance of data to balance the hash bits. In addition, we rotate data using an orthogonal matrix to preserve the topology

structure of data in Hamming space. As a result, multiple hash tables together can balance the precision and recall.

The rest of this paper is organized as follows. Section 2 summarizes the framework of existing multiple hashing tables methods in detail. Section 3 analyzes the PCA-projected data distribution, and then present our method. Section 4 illustrates the advantages of processing data for different hash tables. Experimental results justify the superiority of our method to the existing methods. Section 5 is the conclusions.

2 Related Works

In this section, we review LSH and complementary hashing(CH) methods. Suppose that dataset $X = \{x_i\}_{i=1}^N$ consists N data points, and each of which is a d dimensional vector, that is $x_i \in R^d$. L denotes the number of hash tables. The goal of multiple hash tables method is to learn a binary matrix $B = \{b_i\}^{N \times K}$ for each hash table, where K denotes the length of hashing codes. In general, one hash function maps a data point to a single binary bit $h_k(x) \in \{0, 1\}$. Therefore, there are K hash functions $H = [h_1(x), \dots, h_K(x)]$ for each hash table.

The main goal of LSH[2] is to project the data into a low-dimensional hamming space, and the property of locality will be preserved during the projection to a K bit binary code using K binary-valued hash functions $H = [h_1(x), \dots, h_K(x)]$. In order to preserve the locality property, hash functions $h_k(\cdot)$ of LSH should satisfy:

$$P\{h(x_i) = h(x_j)\} = sim(x_i, x_j) \quad (1)$$

where $sim(x_i, x_j)$ is the similarity function of data x_i and x_j , that is to say, the probability that two point collide in one hash table is equal to their similarity. Different similarity functions result in different kinds of LSH functions, and the classical one is in the form of

$$h(x) = sgn(w^T x + b) \quad (2)$$

where w is a random hyperplane, which is sampled randomly from a p -stable distribution, and b is a random threshold. LSH constructs L hash tables $\{H_l\}_{l=1}^L$ using the above parameters. Therefore, it projects each data point to LK -bit hash codes, and data points with the same hash codes are returned from each table. In order to obtain high precision, K has to be large, which will result in large L to satisfy the collision probability. For large scale application, large L and K will lead to sufficient increase in storage and decelerate the search heavily. Therefore, LSH method works well for long hash codes, but for short codes, they do not present good discrimination[5].

Instead of using random hyperplanes and random thresholds, CH[1] learns parameters of Eqn.(2) from the data. The hashing function of CH is in form of

$$J(\{H_l\}_{l=1}^L) = \sum_{i,j=1}^N \left(a_{ij} \min_{l=1 \dots L} \|H_l(x_i) - H_l(x_j)\|^2 \right) \quad (3)$$

In order to solve this problem, a boosting-based method is adopted, and hashing function is as the classifier. A weight matrix S is defined and updated according to the current classifier H . For the element predicted correctly by current classifier, its weight is set to zero and will not change any more in the future updates, otherwise, the weight would be adjusted to reflect the degree of the contradiction of the original similarity and the similarity in the Hamming space. Given the weighted elements, a hash projection is learned by maximizing the following objective function:

$$\hat{J}(H) = \sum_{i,j=1}^N s_{ij} H(x_i)^T H(x_j) \quad (4)$$

With the constraint relaxed as [23], the objective function is transformed to

$$\hat{J}(H) = \text{tr}[W^T X S X^T W + \eta W^T X X^T W] \quad (5)$$

CH learns hashing functions sequentially in a boosting manner. So that, in advantage of complementarity property of multiple hash tables, the nearest neighbors of query missed from the active bucket of one hash table maybe searched in the next hash table.

3 Weakly Principal Component Hashing

In this section, we present weakly principal component hashing methods, which constructs multiple hash tables and makes a tradeoff between precision and recall rate. The goal of our method is to learn hash functions that preserve the data locality structure in the hamming space, while maximally satisfying the desirable properties of hashing such as independence of bits and partitioning balance.

Firstly, we describe the observation that neighbors which are far in the top K principal component directions maybe near in the next weakly principal component directions. Based on this observation, we project data on different principal component directions to generate inputs for hashing functions learning, which can help to construct complementary hashing functions. Secondly, hashing functions are learned for each hash tables using the generated inputs, in a way that hashing functions learning is transformed to a PCA problem. Finally, in order to optimize the principal component directions, an orthogonal matrix is used to rotate the directions, and improve the precision in the end.

3.1 Choice of Weakly Principal Component Direction

To analyze whether the first K principal directions are sufficient to describe the data, we perform experiment for illustration. Given a set of 256 dimensional data points Y , suppose d_1, \dots, d_K are the top K principal directions. Here, we define the sequential principal directions $d_i, d_i + 1, \dots, d_i + K - 1$ ($i > 1$) as weakly principal component directions. We project Y to different principal component directions as shown in fig.1, where the red 'circle' point is query, other points are its neighborhoods which should be retrieved.

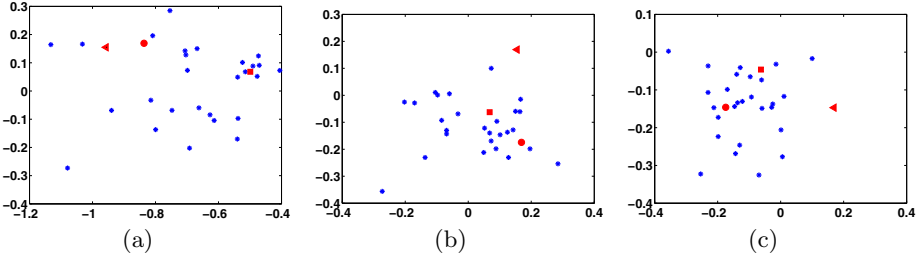


Fig. 1. The distribution of projected data in different principal component directions. (a) projection on d_1, d_2 , (b) projection on d_2, d_3 , (c) projection on d_3, d_4 .

We project the data to d_1, d_2 directions in fig.1(a), and find that the red ‘triangle’ point can be searched and the ‘square’ one would be missed by the small Hamming ball. If omit the first principal direction, and project data to d_2, d_3 as shown in fig.1(b), we find that the the ‘square’ one may be searched with the small Hamming ball. It has the same situation in fig.1(c).

Figure 1 explain why we adopt multiple hashing tables. For the single hashing table, which is the case in fig.1(a), many positive points maybe missed, that is to say, the recall is low. If we adopt two hashing tables, the missed points by one hashing table may be searched in the next hashing table, which can improve the recall. From fig.1, the projected data in different principal directions together can help to search more positive neighbors. If the precision of every hashing table is high, then multiple hashing tables will balance the precision and recall, in a way of improving the recall. Based on the above analysis, we consider these weakly principal components and learn hashing functions, which will increase the number of neighbors and help to retrieve the missing samples, as well as improve the precision.

Let us denote the data matrix as $Y = \{y_i\}_{i=1}^N$, where each column $y_i \in R^d$ is a data point. We assume the data are zero-centered $\sum_{i=1}^N y_i = 0$. Suppose there are L hash tables and K hash functions $\{H_l\}_{l=1}^L$ with $H_l = [h_{1,1}(x), \dots, h_{l,K}(x)]$ for each hash table. $\{X_l\}_{l=1}^L$, $X_l = \{x_{l,i}\}_{i=1}^N \in R^{d \times N}$ be L sets of data for L hash tables respectively. X_l is generated from Y in a way of energy attenuation. The aim is that X_l should include most information of original data, and also can be distinguished for every sample. Based on the analysis above, we extract the eigenvalue $\lambda = \{\lambda_i\}_{i=1}^d$ of the covariance matrix YY^T , and rank them in descending order, and $V = \{v_i\}_{i=1}^d$ is the corresponding eigenvector. In theory, the data projected on the first principal component direction include most energy of the original data, such that, we generate X_l for the l^{th} hash table as follows

$$X_l = Y - (V_l V_l^T) Y \quad (6)$$

where $V_l = \{v_1, \dots, v_l\}$ is the top l eigenvectors of V .

3.2 Hashing Function Construction

For every hash table, each hash function is defined as $h_{l,k}(x) = \text{sgn}(w_{l,k}^T x + b_{l,k})$, which is a linear projection. Since the data is zero-centered, hash function is refined as $h_{l,k}(x) = \text{sgn}(w_{l,k}^T x)$. Excellent hash functions should satisfy the instance that hash codes generated are independent and each bit make a balanced partition of data[6]. Therefore, we learn hash functions by maximizing the following functions.

$$\mathcal{J}(\{H_l\}_{l=1}^L) = \sum_{i=1}^N \max_{l=1 \dots L} \text{var}(H_l(x_{l,i})) = \max_{l=1 \dots L} \sum_{i=1}^N \text{var}(H_l(x_{l,i})) \quad (7)$$

subject to

$$\begin{aligned} \sum_{i=1}^N H_l(x_{l,i}) &= 0, l = 1 \dots L \\ \frac{1}{N} H_l(X_l) H_l^T(X_l) &= I, l = 1 \dots L \end{aligned}$$

By relaxing $H_l(x)$ to its magnitude as well as the constraint like Wang[7], Eqn.(7) is transformed to

$$\begin{aligned} \mathcal{J}\{W_l\}_{l=1}^L &= \frac{1}{N} \text{tr}\{W_l^T X_l X_l^T W_l\} \\ W_l^T W_l &= I \end{aligned} \quad (8)$$

From Eqn.(8), the learning of optimal W_l is transformed to an eigenvector problem. We compute W_l corresponding to the top K maximum eigenvalues of the covariance matrix $X_l X_l^T$.

Optimal hash functions can preserve the original locality structure of the data in the hamming space, which means neighborhoods in the Euclidean space will also be neighborhoods in the hamming space. For the projected data $W_l^T X_l$, quantizing them directly to generate hash codes will result in large error[9]. The quantization loss will be smaller if we rotate W_l using an $K \times K$ orthogonal matrix R_l , in a way of minimizing the following function:

$$\mathcal{Q}(B_l, R_l) = \|B_l - R_l^T W_l^T X_l\|_{\mathbb{F}}^2 \quad (9)$$

where $\|\cdot\|_{\mathbb{F}}$ is the Frobenius norm. An iterative method is adopted to compute B_l and R_l . First, fix R_l and quantize the data points to their nearest vertex of the binary hypercube to obtain B_l . Then, fix B_l to change Eqn.(9) to the classic Orthogonal Procrustes problem and update R_l to minimize the quantization loss. For two points y_i and y_j , their final hamming distance is computed based on the hamming distance of L hash tables.

Algorithm: Weakly Principal Component Hashing with Multiple Tables

Input:data set \mathbf{Y} , length of hash codes \mathbf{K} , number of hash tables \mathbf{L} .**Output:**hash projections $\{H_l\}_{l=1}^L$ 1: Generate X_l for every hash table using (6).2: for $l = 1 : L$ 3: Learn the first top \mathbf{K} projection directions W_l using (8).4: Compute the orthogonal matrix R_l using (9) to rotate W_l .

5: end for

6: Output $H_l = R_l^T W_l^T X$.

4 Experiments

4.1 Image Dataset and Evaluation Items

In this section, we evaluate the proposed method(named WPCH) on CIFAR dataset[9][23] and SIFT-1M dataset[26]. We compare WPCH method with 6 state-of-the-art binary coding methods, which are ItQH, unsupervised SPLH (USPLH), CH, SH, LSH and Anchor Hashing(AH).

We evaluate these methods under two schemes:1)F-measure: for the positive points among the retrieved samples whose hash codes fall within a Hamming radius r around the query's hash code, we compute the precision and recall of each case, and then use F-measure to evaluate different methods. 2)Hamming ranking: the retrieved points are ranked according to the corresponding Hamming distance, and the precision of top N images is recorded. In comparison experiments, the multiple hash tables methods WPCH and CH are both with $L = 3$.

4.2 Results on CIFAR Dataset

CIFAR dataset is sampled from labeled subsets of the 80 million tiny images, and has been used as a benchmark dataset for designing hash code methods[9][23]. A subset of 1,000 data points are randomly selected to construct the test set and the other samples are taken as the training database to learn hash functions. For every test sample, its ground truth neighbors are those whose Euclidean distance are within the top 5% of the whole database. Figure 2 shows the mean average precision(mAP) results for WPCH with different number of hashing tables L . Compared to $L = 1$, the mAP of WPCH with $L = 2$ is much higher, and then achieves the maximum for $L = 3$. For hash tables $L > 3$, the mAP decreases slowly if the number of hash tables continues increasing. That is because giving up the principal component direction can improve the distinction of projected data. However, More principal component information omitted will reduce the data energy roughly, as a result, the projected data cannot preserve the information of the original data well.

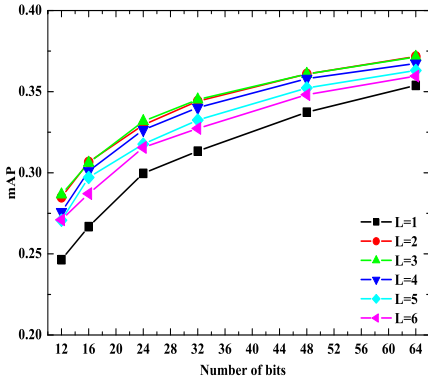


Fig. 2. Performance of mAP for different number of hashing tables of WPCH on CIFAR dataset

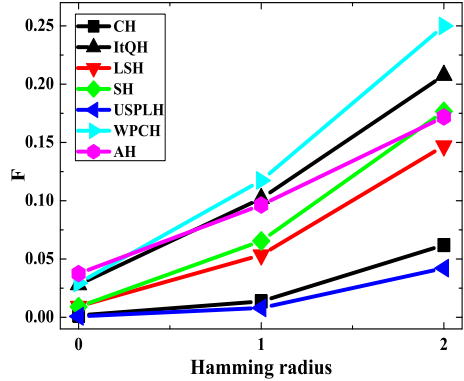


Fig. 3. Performance of F-measure for different different methods on CIFAR dataset. The three points in every curves corresponding hamming radius $r = 0$, $r = 1$, $r = 2$ respectively.

In practice, if the precision is high, the recall is low, and vice versa. It is difficult for them to reach high at the same time. So, we use F-measure to reflect the overall performance. Figure 3 shows the F-measure of 16 bits Hamming radius $r = 0, 1, 2$ for different methods on CIFAR dataset. In fig.3, the F-measure of WPCH with $L = 3$ is a little lower than AH for $r = 0$ case. The F-measure of WPCH with $L = 3$ is increased with the radius increasing. For $r = 1, 2$ instances, the F-measure of WPCH with $L = 3$ is higher than other methods.

Figure 4 illustrates the performance of different methods using Hamming ranking. For the large-scale application, our method still works better than the other methods. For the other two multiple hash tables methods, LSH is worse for short code, and CH is worse for long code.

4.3 Results on SIFT-1M Dataset

SIFT-1M dataset is a large scale dataset to evaluate the quality of approximate nearest neighbors search algorithm. It contains 1 million data points, which are represented by 128- d SIFT descriptors.

Figure 5 is the mAP of WPCH with different hash tables on SIFT-1M dataset. Compared with $L = 1$, the mAP of WPCH with $L = 3$ is improved 15% in the case of 12 bits. For other code bits, the precision of multiple hash tables all exceed single table instance. From Fig.5, the effective of adopting multiple hash tables is obvious especially for short hashing codes(i.e. 12, 16, 24 and 32 bits). As the same with CIFAR dataset, the performance of mAP is better for multiple hash tables than single table on large scale dataset.

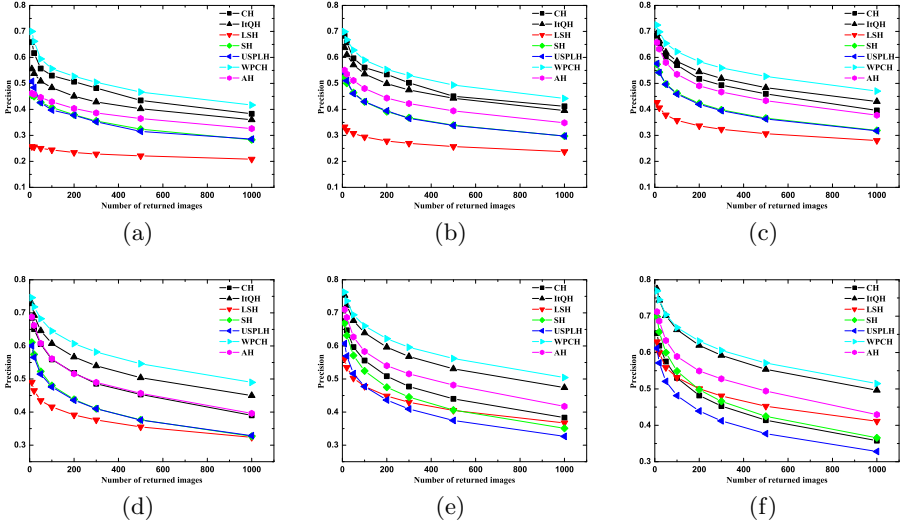


Fig. 4. On CIFAR dataset, comparison of the performance using Hamming ranking. (a)-(f) are the performance for the hash codes of 12, 16, 24, 32, 48 and 64bits respectively.

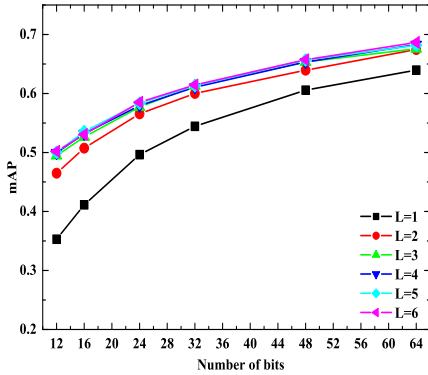


Fig. 5. Performance of mAP for different bits of WPCH on SIFT-1M dataset

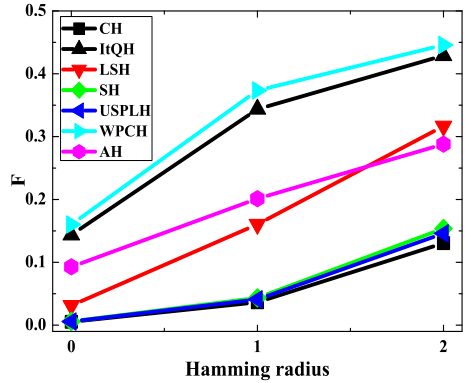


Fig. 6. Performance of F-measure for different different methods on SIFT-1M dataset. The three points in every curves corresponding hamming radius $r = 0$, $r = 1$, $r = 2$ respectively.

Figure 6 shows the F-measure of 16 bits Hamming radius $r = 0, 1, 2$ for different methods on SIFT-1M dataset. the F-measure of $WPCH(L = 3)$ exceeds other methods. The proposed method is excellent for all instances.

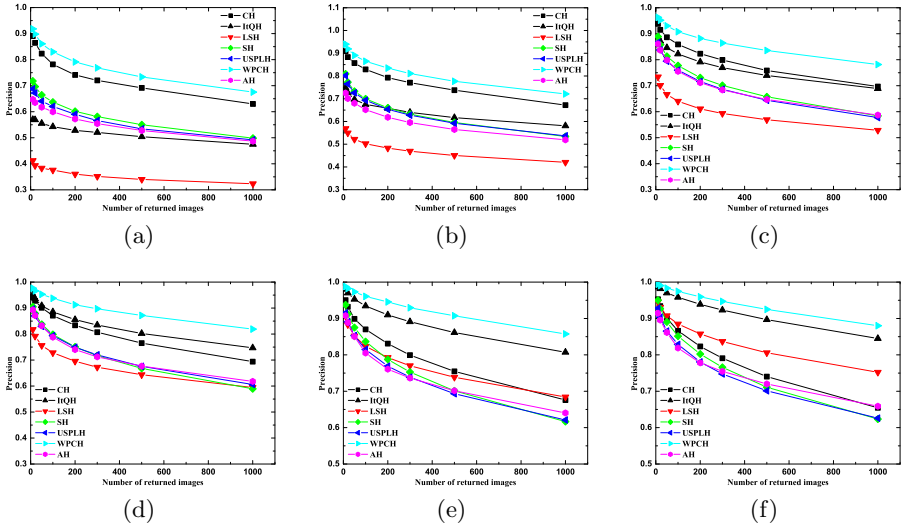


Fig. 7. On SIFT-1M dataset, comparison of the performance using Hamming ranking. (a)-(f) are the performance for the hash codes of 12, 16, 24, 32, 48 and 64 bits respectively

Hashing rank is used to evaluate different hashing method on this dataset. We change the number of returned image, and compute the precision for different hashing methods. Figure 7 shows comparison results. Fig.7(a)-(d) are the performance of different hashing method with short hashing codes. In these instances, WPCH($L = 3$) is excellent. The performance of CH($L = 3$) and ItQH is following. For long hashing codes(i.e. codes longer than 32 bits), the precision of our method is a little higher than ItQH method, and much higher than other methods.

5 Conclusion

In this paper, we propose an unsupervised hashing method, which learns binary codes in a data-blind way and constructs hashing functions with multiple hash tables to balance the precision and recall rate. For each hash table, we process data points to balance the data energy and detail, which can increase the distinction among points for Hamming projection. We experimentally illustrate the superior performance of the proposed method to existing ones on three large scale image datasets.

Acknowledgments. This paper is supported by Major Program of National Natural Science Foundation of China (No. 70890080 and 70890083).

References

1. Xu, H., Wang, J., Li, Z., Zeng, G., Li, S., Yu, N.: Complementary Hashing for Approximate Nearest Neighbor Search. In: 24th IEEE International Conference on Computer Vision, pp. 1631–1638 (2011)
2. Andoni, A., Indyk, P.: Near-optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions. *Commun. ACM.* 51, 117–122 (2008)
3. Kulis, B., Grauman, K.: Kernelized Locality-sensitive Hashing for Scalable Image Search. In: 12nd IEEE International Conference on Computer Vision, pp. 2130–2137 (2009)
4. Raginsky, M., Lazebnik, S.: Locality-sensitive Binary Codes from Shift-invariant Kernels. In: 23rd Annual Conference on Neural Information Processing Systems, pp. 1509–1517 (2009)
5. Torralba, A., Fergus, R., Weiss, Y.: Small Codes and Large Image Databases for Recognition. In: 21st IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
6. Weiss, Y., Torralba, A., Fergus, R.: Spectral Hashing. In: 22nd Annual Conference on Neural Information Processing Systems, pp. 1753–1760 (2008)
7. Wang, J., Kumar, S., Chang, S.: Semi-Supervised Hashing for Scalable Image Retrieval. In: 23th IEEE Conference on Computer Vision and Pattern Recognition, pp. 3424–3431 (2010)
8. Wang, J., Kumar, S., Chang, S.: Sequential Projection Learning for Hashing with Compact Codes. In: 27th International Conference on Machine Learning, pp. 1127–1134 (2010)
9. Gong, Y., Lazebnik, S.: Iterative Quantization: A Procrustean Approach to Learning Binary Codes. In: 24th IEEE Conference on Computer Vision and Pattern Recognition, pp. 817–824 (2011)
10. Mu, Y., Chen, X., Chua, T., Yan, S.: Learning Reconfigurable Hashing for Diverse Semantics. In: 1st ACM International Conference on Multimedia Retrieval, pp. 1–8 (2011)
11. Shao, J., Wu, F., Ouyang, C., Zhang, X.: Sparse Spectral Hashing. *Pattern Recogn. Lett.* 33, 271–277 (2011)
12. Chen, C., Horng, S., Huang, C.: Locality Sensitive Hashing for Sampling-based Algorithms in Association Rule Mining. *Expert Syst. Appl.* 38, 12388–12397 (2011)
13. Strecha, C., Bronstein, A.M., Bronstein, M.M., Fua, P.: LDAHash: Improved Matching with Smaller Descriptors. *IEEE T. Pattern Anal* 34, 66–78 (2012)
14. Jagadish, H.V., Ooi, B., Tan, K., Yu, C., Zhang, R.: iDistance: An Adaptive B+-tree Based Indexing Method for Nearest Neighbor Search. *ACM T. Database Syst.* 30, 364–397 (2005)
15. Yang, K., Shahabi, C.: An efficient k nearest neighbor search for multivariate time series. *Inform Comput.* 205, 65–98 (2007)
16. Giannella, C.: New instability results for high-dimensional nearest neighbor search. *Inform Process. Lett.* 109, 1109–1113 (2009)
17. Brandt, J.: Transform coding for fast approximate nearest neighbor search in high dimensions. In: 23th IEEE Conference on Computer Vision and Pattern Recognition, pp. 1815–1822 (2010)
18. Liu, D.: A strong lower bound for approximate nearest neighbor searching. *Inform Process. Lett.* 92, 23–29 (2004)
19. Cha, G., Zhu, X., Petkovic, D., Chung, C.: An efficient indexing method for nearest neighbor searches in high-dimensional image databases. *IEEE T. Multimedia* 4, 76–87 (2002)

20. Indyk, P.: Nearest-neighbor-preserving embeddings. *ACM Transactions on Algorithms* 3, 1549–6325 (2007)
21. Yu, D., Zhang, A.: ClusterTree: integration of cluster representation and nearest-neighbor search for large data sets with high dimensions. *IEEE T Knowl. Data En.* 15, 1316–1337 (2003)
22. Jégou, H., Douze, M., Schmid, C.: Improving Bag-of-Features for Large Scale Image Search. *Int. J. Comput Vision* 87, 316–336 (2010)
23. Wang, J.: Semi-Supervised Learning for Scalable and Robust Visual Search. PhD. Thesis, Columbia University, USA (2011)
24. Zhou, J., Fu, H., Kong, X.: A balanced semi-supervised hashing method for CBIR. In: 18th IEEE International Conference on Image Processing, pp. 2481–2484. IEEE Press, New York (2011)
25. Chua, T., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In: 8th ACM International Conference on Image and Video Retrieval. ACM Press (2009)
26. Jégou, H., Douze, M., Schmid, C.: Product Quantization for Nearest Neighbor Search. *IEEE T. Pattern Anal.* 33(1), 117–128 (2011)
27. Liu, W., Wang, J., Kumar, S., Chang, S.: Hashing with Graphs. In: 28th International Conference on Machine Learning (2011)
28. Wang, M., Yang, K., Hua, X., Zhang, H.: Towards a Relevant and Diverse Search of Social Images. *IEEE T. Multimedia* 12(8), 829–842 (2010)
29. Wang, M., Hua, X., Tang, J., Hong, R.: Beyond Distance Measurement: Constructing Neighborhood Similarity for Video Annotation. *IEEE T. Multimedia* 11(3), 465–476 (2009)

DUT-WEBV: A Benchmark Dataset for Performance Evaluation of Tag Localization for Web Video

Haojie Li, Lei Yi, Yue Guan, and Hao Zhang

School of Software, Dalian University of Technology
hjli@dlut.edu.cn, yilei@mail.dlut.edu.cn, worm004@hotmail.com

Abstract. Nowadays, numerous social videos have pervaded on the Web. Social web videos are characterized with the accompanying rich contextual information which describe the content of videos and thus greatly facilitate video search and browsing. Generally those context data such as tags are generated for the whole video, without temporal indication on when they actually appear in the video. However, many tags only describe parts of the video content. Therefore, tag localization, the process of assigning tags to the underlying relevant video segments or frames is gaining increasing research interests and a benchmark dataset for the fair evaluation of tag localization algorithms is highly desirable. In this paper, we describe and release a dataset called *DUT-WEBV*, which contains 1550 videos collected from *YouTube* portal by issuing 31 concepts as queries. These concepts cover a wide range of semantic aspects including scenes like “mountain”, events like “flood”, objects like “cows”, sites like “gas station”, and activities like “handshaking”, offering great challenges to the tag (i.e., concept) localization task. For each video of a tag, we carefully annotate the time durations when the tag appears in the video. Besides the video itself, the contextual information, such as thumbnail images, titles, and categories, is also provided. Together with this benchmark dataset, we present a baseline for tag localization using multiple instance learning approach. Finally, we discuss some open research issues for tag localization in web videos.

Keywords: Video annotation, tag localization, video retrieval.

1 Introduction

With the popularity of web 2.0, recent years have witnessed the proliferation of social videos and the success of many social websites, including *YouTube*, *MySpace*, etc. These websites not only allow users to upload and share their generated videos, but also encourage them to describe the content of videos with context information like tag, category, title, and so on. Such information greatly facilitates video retrieval, browsing, and sharing by using text based search method [1]. Generally those context data such as tags are generated for the whole video, saying, they are video-level annotations, while many tags are

actually only related to parts of the video content. Thus the performance of tag based video search is reduced. For example, a video of sports news can be tagged with “basketball”, though only a few shots or frames are actually talking about “basketball”. When we use “basketball” as keyword to search, this video may be returned to us. In such scenario, we need watch the whole video to find parts of interests, which is quite boring and time consuming. Thus for better user experiencing and more precise video searching, it is demanding to know what time durations or points of the video are actually related to a tag, *i.e.*, we need localize tags to related video shots or frames.

Recently, tag localization of web videos is becoming the key to a wide variety of multimedia applications like training concept models [2][3], precise video retrieval [4], web video browsing and summarization [5][6], etc, and more and more research efforts have been dedicated to tackle this problem. In [2], Ulges *et al.* learnt concept detectors from weakly labeled web videos. To reduce the influences of noisy irrelevant contents on classifier performance, they proposed to localize the relevant shots to a given concept by modeling the relevance of video frames as a latent random variable in a probabilistic framework and estimated the latent variables in an EM fashion. Ballan *et al.* [7][8] proposed to utilize the social knowledge to suggest and localize tags in web videos. They first retrieved *Flickr* images using video tags and selected visual similar images for keyframes of shots, then the ranked retrieved tags of *Flickr* images are suggested for each shot. Similarly, Chu *et al.* [9] used *Flickr* images and the associated tags for tag localization, while modeled relationship between keyframes in a video shot and candidate tags as a bipartite graph and the best matching was found to determine the most appropriate tags. In [4], Guangda *et al.* assigned tags to shot-level segments of a video in an improved multiple instance learning framework by considering tag correlations and temporal smoothness. In [5], they further used this technique to identify key-shots for event driven web video summarization. Based on the assumption that a positive bag may contain arbitrary number of positive samples, Ulges *et al.* [10] proposed a relaxed multiple instance learning strategy to identify relevant frames in video.

Though extensive works have been conducted recently for tag localization in web video, the proposed algorithms were all tested on their own datasets, which prevents the fair comparison of different approaches. Thus, it is highly desirable to present a public benchmark dataset for evaluating and promoting the tag localization research. In this paper, we describe and release such a dataset called *DUT-WEBV*, which is consisted of 1550 videos collected from video sharing website *YouTube* by issuing 31 concepts as queries. These concepts cover a wide range of semantic levels including scenes like “mountain”, events like “flood”, objects like “cows”, sites like “gas station”, and activities like “handshaking”, offering great challenges to the tag (*i.e.*, concept) localization task. In *DUT-WEBV*, for each video of a tag, we carefully annotate the time durations when the tag appears in the video. Besides the videos and the frame-level annotations, the contextual information, such as thumbnail images, titles, and categories, is also provided.

The rest of the paper is organized as follows. In Section 2 we will give a detailed description of *DUT-WEBV* and the comparison with other related datasets. The annotation process is introduced in Section 3. In Section 4 we will present a baseline for tag localization using multiple instance learning approach. Finally, we will discuss the open research issues in Section 5 and conclude in Section 6.

2 *DUT-WEBV* Dataset

To construct the dataset, the first step is to design the concept/tag set that are suitable for testing tag localization algorithms. *LSCOM* [11] is a widely used visual lexicon of more than 834 concepts defined for a variety of research tasks, such as multimedia learning, information retrieval, computational linguistics, etc., according to the criteria of *usefulness*, *feasibility*, and *observability*. We carefully select 30 concepts from *LSCOM* based on two additional criteria of *coverage* and *detestability* as the tag set of *DUT-WEBV*. *Coverage* here means the selected tags should cover a wide range of semantic levels while *detestability* means that the tags should not be too abstract, *i.e.*, they could be detected

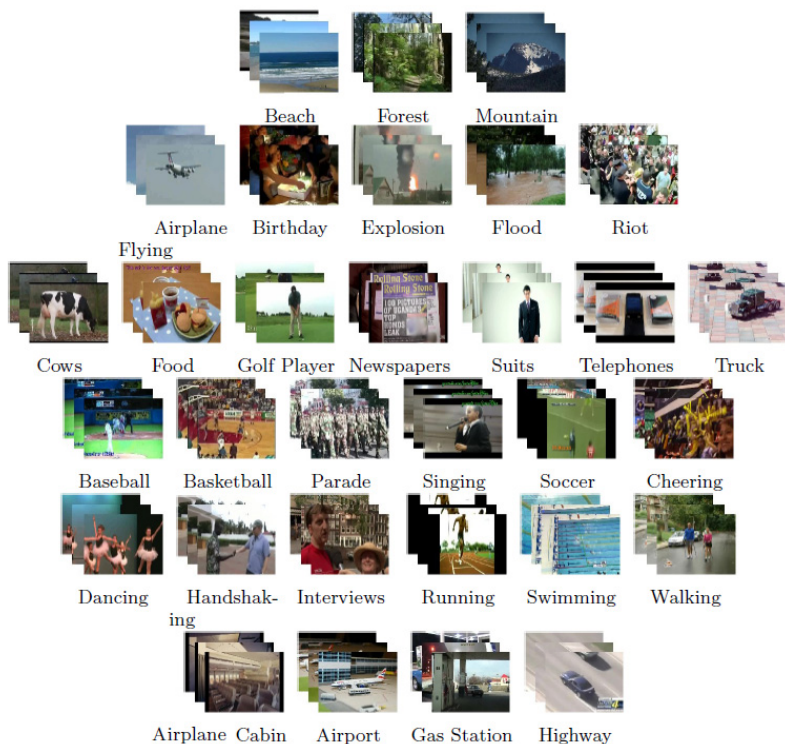


Fig. 1. The concepts(tags) in *DUT-WEBV* and their example videos. The concepts in the first row, second row, third row, fourth & fifth row, and sixth row are **scenes**, **events**, **objects**, **people activities** and **sites** respectively.

based on the visual and/or audio features. One more concept, “birthday”, is selected from *CCV* dataset [12]. The final selected tags/concepts in *DUT-WEBV* are related to objects, events, people activities and scenes, which brings great challenges to the tag localization task. The 31 tags and the example videos are shown in Fig. 1.

The 31 concepts are then used as queries to download videos (video duration is limited to 10 minutes) from *YouTube* to construct the video set for each concept. To ensure the diversity of topic for videos of each tag, we retrieved about 1000 videos from different categories/channels and different users for a given tag. Then we randomly selected videos and manually checked their relevance to the tag and the relevant ones are retained while irrelevant ones are filtered out. Tutorial videos and professionally edited videos are removed since they are not common in real life. In this way, we collected 50 videos for each tag and the final dataset has 1550 videos with an average duration of around 228 seconds. For each video of a tag, we manually annotated the time durations when the tag actually appears in the video using a developed annotation tool (the annotation process will be provided in the next section). Besides the videos and the frame-level annotations, the contextual information which may be useful to design advanced tag localization methods, such as thumbnail images, video urls, titles, users, and categories etc, is also included in this dataset.

2.1 Tag Relevance of *DUT-WEBV*

With the fine annotations on frame level, we can estimate a statistics, relevance, of the dataset to measure how many frames in a video are indeed related to a given tag. The relevance for each tag, category and the whole dataset are shown in Tab. 1 (in Section 4). It can be seen that the average relevance is about 47.8% (Note that we have filtered out irrelevant videos before constructing the dataset), which implies that about 52% of the video frames are actually not related to the given tag. Thus if we directly use web videos as training corpus to construct concept detectors, the mixed noisy content will definitely adversely affect the performance of trained detectors [2].

2.2 Comparison with Related Datasets

There are many video datasets that are related to *DUT-WEBV* and their differences to *DUT-WEBV* will be briefly discussed in the following.

Columbia Consumer Video (CCV) Dataset [12] is collected from *YouTube* for video content understanding research task, especially for interests of consumer area. It contains 9,317 web videos over 20 semantic categories, including events, scenes, and objects. The ground truth is obtained from Internet annotators through the Amazon MTurk platform. Though the content of *CCV* is as diverse as *DUT-WEBV*, it is mainly designed for the performance evaluation of video category classification methods, thus the label is only at the video level.

MCG-WEBV video Dataset [13] is a benchmark dataset for web video analysis, such as topic discovery and track, web video categorization etc. The dataset includes 80,031 most viewed videos for every month from Dec. 2008 to Feb. 2009 on *YouTube* and offers the ground-truth of 73 hot web topics from human annotation.

YouTube 22 Concepts Dataset [14] is constructed for experiments with automatic video annotation, which consists of 2200 real-world online video clips downloaded from *YouTube* for 22 semantic concepts. The ground truth tags given by *YouTube* users are also provided. Similar to *CCV*, the annotations of the last two datasets are all generated for whole video clip, while the temporal location of tag is not provided.

TRECVID SIN 2012 Dataset [15] is currently the largest video dataset in collection size and diversity, containing manual annotations for 346 concepts on around 393,600 keyframes extracted from 600 hours web videos. The aim of this dataset is to provide an open training and evaluation corpus for testing various algorithms of visual concept detectors, thus the annotation granularity is same to *DUT-WEBV*, namely, they are frame-level annotations. However, the annotated keyframes of TRECVID videos are selected randomly through the online collaborative annotation system, thus it is not guaranteed that all the keyframes of a video be annotated. Therefore, TRECVID SIN corpus is not suitable for the evaluation of tag localization algorithms.

3 The Annotation Method

Compared to the category or topic annotation of video at video level, temporally localizing the range of a tag in video is really time-consuming. In this work, we designed and developed an annotation tool to help annotators efficiently specify the time durations of tag in video. This tool contains three main interfaces: concept and unlabeled video selection interface (*CUVS*), relevant keyframe selection interface (*RKS*) and precise time period selection interface (*PTPS*).

In *CUVS* interface, the annotator can select a concept to annotate from a list box where all the 31 concepts and their annotation status are listed. Then the annotator is allowed to select an unlabeled video randomly or by specifying the folder. After selecting concept and videos, the *RKS* interface will be activated as shown in Fig. 2. The tool extracts one keyframe per second from the video and a group of 12 keyframes are displayed in the right panel of *RKS*. This interface allows the annotator to glance at the panel to quickly and coarsely find relevant keyframes. There are two choices for each keyframe, i.e., *yes* and *no* indicating the keyframe is relevant to the tag or not (the default choices for all keyframes are *no*). Related keyframes will be marked with red boxes. An enlarged picture of the keyframe is shown in the upper part of left panel when the mouse slides on the keyframe. The detailed information on the annotated video is displayed in the bottom part of left panel.

Since keyframe is evenly extracted at each second, keyframe-based annotation is too coarse for evaluation of tag localization algorithms and fine-granularity

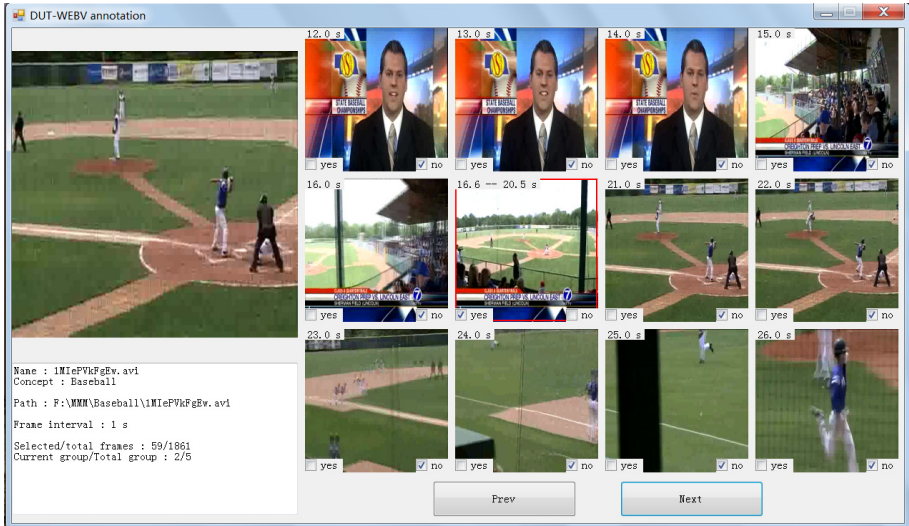


Fig. 2. Interface for relevant keyframe selection

annotation on when the given tag appears in the video is needed. By right clicking on the selected relevant keyframe, the *PTPS* interface (as shown in Fig. 3) will pop-up where a set of buttons are configured to allow the annotator to easily specify the precise time duration when the tag appears. We take Fig. 2 and Fig. 3 as example to explain the precise annotation process. In Fig. 2 we observed that the third picture in the second row (the keyframe for the 21st second) is related to the concept “golf player”, and we can right click the picture, then Fig. 3 is popped up. In Fig. 3, a video player is embedded in which the video segment from 20th second (i.e., 21 plus one) to the end of video is presented for judging, based on the assumption that the following content of current keyframe will be very likely to be relevant. A sliding bar is provided for determining the start point and ending point of relevant segment with buttons “{” and “}” respectively. The selected segment can also be finely tuned using the provided *forward* and *backward* buttons. After selecting the precise time duration, the keyframes inside this time duration will disappear from Fig. 2 and the clicked picture will be replaced by the frame of the endpoint. The time in the top left region will also be replaced by the time duration. In this example, the frames of time period from 16.6 second to 20.5 second have already been specified as related to concept “golf player”.

4 Tag Localization Baseline

We present a tag localization baseline on the dataset using multiple instance learning (*MIL*) techniques. *MIL* has been successfully applied to image annotation [16], tag location for Internet videos [4] and relevant frames identification

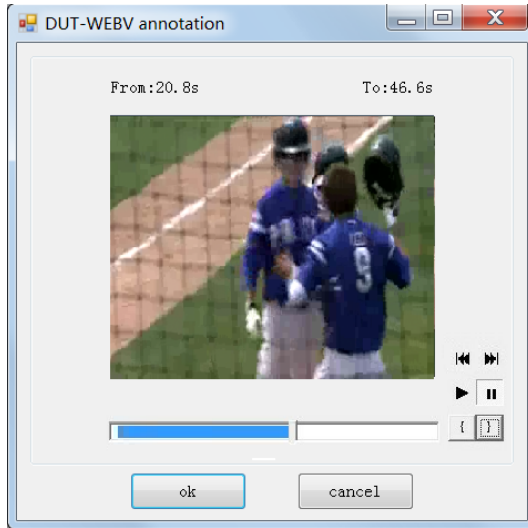


Fig. 3. Interface for precise time duration selection

in weakly labeled videos [2]. In *MIL*, labels are assigned to bags of instances instead of single instance. A bag is composed of multiple instances and, a bag is labeled positive if any of its instances is positive; a bag is labeled negative when all of its instances are negative. Given the bag label, many approaches have been developed to infer the instance label. Thus if we treat a video as bag and the frames as instances, the tag localization task can be naturally modeled as a *MIL* problem. In this work, we adopt the MIL-BPNET approach proposed in [17] to localize the relevant parts in a positive video to a certain tag. Given a tag, the videos of the tag form the positive bags while videos from other tags form the negative bags. To reduce computational cost, for each tag, we manually select 10 negative tags and then randomly sample 10 videos from each negative tag to build the negative bags of the tag.

In the baseline method, for simplicity, we sample one frame of every two seconds from the videos and then estimate the relevance of these frames to given tags. For feature representation, we extract the 128 dimensional *SIFT* (Scale-Invariant-Feature-Transformation) [18] feature of video frames and quantized them into 500 dimensional bag of word (*BoW*) representation using the hierarchical clustering method [19]. After *MIL*, each sampled frame is associated with a relevance score to its corresponding tag and it will be labeled as positive if the score exceeds a threshold, otherwise labeled as negative. However, it is difficult to determine the threshold for each tag, thus the *precision* metric is not suitable to evaluate the performance of tag localization method. Here we propose to use $P@N$ (precision at top N ranked results) and improved $P@N$ ($iP@N$) as the measurements, while N is adaptively determined according to the relevance statistic of each tag as follows.

$$N = C * r$$

where C is the total positive frames of the tag and r is the relevance rate. The motivation is that, since we have the relevance rate r for a tag t , thus if we select the top $N = C * r$ ranked results belonging to tag t after *MIL*, the precision, $P@N$, should be larger than the precision of randomly selected N samples (*i.e.*, r). Let $iP@N$ be the improvement of $P@N$ against r , then the larger $iP@N$ is, the better performance is achieved by the baseline method. The parameters for MIL-BPNET training are set to 20 hidden units, 200 training epochs and a fixed back propagation learning rate (0.05). The results for the 31 tags are tabulated in Tab. 1.

From Tab. 1 we can observe that, by applying tag localization via *MIL* we have an average of relevance improvement of 18.9% @ N , showing the effectiveness of *MIL* in localizing tag to frames. However, the improvements on different tags/categories are not equivalent, ranging from 12.3% to 30.4%, which reveals two facts: 1) the visual (*SIFT* feature here) detestability of different tags/categories is distinct; 2) compared with the baseline method, there is much room for improvement for tag localization in videos.

5 Open Issues

Due to its importance in various multimedia applications, tag localization of web videos is attracting increasing research attentions in recent years. However, the research of tag localization is still in its early stage and many research issues are open.

1) **Novel machine learning algorithms are demanding.** In many previous works [4][10], tag localization in videos is treated as a *MIL* task where the video is regarded as bag and keyframe or shot is regarded as instance. Thus tag localization is equivalent to identifying the positive instances in the positive bag. However, *MIL* is based on the assumption that a positive bag must have at least one positive instance. As pointed out in Section 2, this assumption does not always hold for web videos. The relevance of a tag to a video is rather complex, *i.e.*, a video tagged with a tag may have none, a few or many relevant frames. In [2], to find the relevant frames using KDE method, Ulges *et al.* take the relevance as a prior. It is, however, difficult to obtain the prior for any video and any tag. Therefore, to handle the complexity of tag relevance to video, novel advanced machine learning algorithms need be investigated and introduced to the tag localization in web videos.

2) **Multi-modality fusion for tag localization.** Current works on web video tag localization are actually image-based methods, where only visual features of video keyframes are used. However, most concepts or tags are multimodal in nature [20]. While some tags are primarily expressed by the visual aspect (e.g., beach, mountain) or audio aspect (e.g., singing), or characterized by motion cues (e.g., running, walking), most tags involve more than one aspects of the visual, audio and motion information simultaneously. Thus, to boost the performance of tag localization systems, some research issues maybe: how to select the ap-

Table 1. The relevance statistic and results of baseline method for each tag and category on the dataset

Category	Concept/Tag	Relevance	P@N	iP@N
Events(5)	airplane flying	0.394	0.726	0.842
	birthday	0.272	0.305	0.121
	explosion	0.505	0.650	0.287
	flood	0.524	0.550	0.050
	riot	0.671	0.693	0.033
		0.473	0.585	0.267
Objects(7)	cows	0.578	0.581	0.005
	food	0.253	0.416	0.644
	golf player	0.324	0.386	0.191
	newspapers	0.381	0.416	0.092
	suits	0.399	0.425	0.065
	telephones	0.472	0.534	0.131
	truck	0.357	0.521	0.460
		0.395	0.468	0.227
People activities (12)	baseball	0.607	0.669	0.102
	basketball	0.591	0.643	0.088
	cheering	0.408	0.582	0.426
	dancing	0.272	0.281	0.033
	handshaking	0.429	0.447	0.041
	interviews	0.574	0.618	0.076
	parade	0.585	0.694	0.186
	running	0.450	0.455	0.011
	singing	0.580	0.611	0.053
	soccer	0.667	0.763	0.144
	swimming	0.549	0.708	0.290
	walking	0.419	0.430	0.026
		0.532	0.575	0.123
Scene(3)	beach	0.549	0.705	0.284
	forest	0.594	0.732	0.232
	mountain	0.411	0.574	0.397
		0.518	0.670	0.304
Sites(4)	aircraft cabin	0.495	0.519	0.048
	airport	0.631	0.701	0.111
	gas station	0.222	0.235	0.059
	highway	0.438	0.585	0.336
		0.447	0.510	0.139
Total		0.478	0.553	0.189

preciate modality for certain tags, and how to effectively and efficiently fuse multi-modality features.

3) **Context mining for tag localization.** Social web videos are accompanied with rich contextual information, including thumbnail images, titles, categories, and user ids, etc. Such information can be mined and leveraged for tag localization task since videos with same context (e.g., videos coming from same

user or same category) usually have same content. On the other hand, videos are generally tagged with a few tags and these tags have high semantic correlations. For example, a video shot or keyframe tagged with “fire” will have a high probability of being tagged with “explosion”, therefore, based on this observation, if a shot or keyframe is localized by “fire”, we can increase its confidence of localization of “explosion”. Actually, such kind of concept co-occurrence relationship has been widely exploited in image annotation and retrieval [21] and demonstrated its effectiveness, while having not been investigated in tag localization before. We expect the mining of the rich context of videos will benefit tag localization.

6 Conclusions

In this paper, we described a benchmark dataset crawled from *YouTube* for the evaluation of tag localization algorithms of web videos. The collecting of tags and videos, and the annotation process are detailed presented. For the connivance of performance comparison, a baseline method based on multiple instance learning is also provided. The context information of videos is included in the dataset to allow for development of more advanced algorithms. In the future, we will extend the dataset in size of concept/tag number and video number per tag.

Acknowledgements. This work was supported by National Natural Science Funds of China (61033012, 61173104).

References

1. Wang, M., Ni, B., Hua, X.-S., Chua, T.-S.: Assistive Tagging: A Survey of Multimedia Tagging with Human-Computer Joint Exploration. *ACM Computing Surveys* 44(4) (2012)
2. Ulges, A., Schulze, C., Breuel, T.: Identifying Relevant Frames in Weakly Labeled Videos for Training Concept Detectors. In: *ACM CIVR* (2008)
3. Ikizler-Cinbis, N., Cinbis, R.G., Sclaroff, S.: Learning Actions From the Web. In: *International Conference on Computer Vision* (2009)
4. Li, G., Wang, M., Zheng, Y.-T., Li, H., Zha, Z.-J., Chua, T.-S.: ShotTagger: tag location for internet videos. In: *ICMR* (2011)
5. Wang, M., Hong, R., Li, G., Yan, S., Chua, T.-S.: Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification. *IEEE Trans. on Multimedia* 14(4), 975–985 (2012)
6. Hong, R., Tang, J., Tan, H.-K., Ngo, C.-W., Yan, S., Chua, T.-S.: Beyond search: Event-driven summarization for web videos. *TOMCCAP* 7(4), 35 (2011)
7. Ballan, L., Bertini, M., Del Bimbo, A., et al.: Tag suggestion and localization in user-generated videos based on social knowledge. In: *Proc. of the 2nd ACM SIGMM International Workshop on Social Media* (2010)
8. Ballan, L., Bertini, M., Del Bimbo, A., Serra, G.: Enriching and localizing semantic tags in internet videos. *ACM Multimedia* (2011)
9. Chu, W.-T., Li, C.-J., Chou, Y.-K.: Tag suggestion and localization for web videos by bipartite graph matching. In: *Proc. of the 3rd ACM SIGMM International Workshop on Social Media, WSM 2011* (2011)

10. Ulges, A., Schulze, C., Breuel, T.: Multiple Instance Learning from Weakly Labeled Videos. In: SAMT Workshop on Cross-Media Information Analysis and Retrieval (2008)
11. Naphade, M., Smith, J.R., Tesic, J., Chang, S.-F., Hsu, W., Kennedy, L., Hauptmann, A., Curtis, J.: Large-Scale Concept Ontology for Multimedia. *IEEE Multimedia* 13, 86–91 (2006)
12. Jiang, Y.-G., Ye, G., Chang, S.-F., Ellis, D.P.W., Loui, A.C.: Consumer video understanding: a benchmark database and an evaluation of human and machine performance. In: *ICMR* (2011)
13. Cao, J., Zhang, Y.D., Song, Y.C., Chen, Z.N., Zhang, X., Li, J.T.: MCG-WEBV: A Benchmark Dataset for Web Video Analysis. Technical Report, ICT-MCG-09-001, Institute of Computing Technology (May 2009)
14. Ulges, A., Schulze, C., Keysers, D., Breuel, T.M.: A System That Learns to Tag Videos by Watching Youtube. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds.) *ICVS 2008*. LNCS, vol. 5008, pp. 415–424. Springer, Heidelberg (2008)
15. <http://www-nlpir.nist.gov/projects/tv2012/tv2012.html>
16. Tang, J., Li, H., Qi, G.-J., Chua, T.-S.: Image Annotation by Graph-Based Inference With Integrated Multiple/Single Instance Representations. *IEEE Transactions on Multimedia* 12(2), 131–141 (2010)
17. Zhang, M.-L., Zhou, Z.-H.: Improve Multi-Instance Neural Networks through Feature Selection. *Neural Process Letters* 19(1), 1–10 (2004)
18. Tang, S., Zheng, Y.-T., Wang, Y., Chua, T.-S.: Sparse Ensemble Learning for Concept Detection. *IEEE Transactions on Multimedia* 14(1), 43–54 (2012)
19. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: *CVPR* (2006)
20. Shen, J., Tao, D., Li, X.: Modality Mixture Projections for Semantic Video Event Detection. *IEEE Trans. Circuits Syst. Video Techn.* 18(11), 1587–1596 (2008)
21. Wang, M., Yang, K., Hua, X.-S., Zhang, H.-J.: Towards a Relevant and Diverse Search of Social Images. *IEEE Transactions on Multimedia* 12(8), 829–842 (2010)

Clothing Extraction by Coarse Region Localization and Fine Foreground/Background Estimation

Xiao Wu, Bo Zhao, Ling-Ling Liang, and Qiang Peng

Department of Computer Science and Engineering, Southwest Jiaotong University
No. 111, North Section 1, 2nd Ring Road, Chengdu, China
{wuxiaohk, qpeng}@home.swjtu.edu.cn,
{zhaobo1987, eaily471956454}@gmail.com

Abstract. Online shopping is becoming more and more popular for billions of web users because of its convenience and efficiency. Customers can use content-based product image search engine to find their desired products. However, a frustrating fact is that the search results are significantly affected by the presence of natural backgrounds and fashion models. To minimize the influence of these noises, in this paper, an automatic clothing extraction algorithm is proposed, which consists of two phases: coarse clothing region localization with human proportion, and fine foreground/background modeling. Experiments on two datasets crawled from e-commerce websites demonstrate that the proposed approach achieves good performance, and has competitive performance with the interactive solution.

Keywords: Clothing Segmentation, Gaussian Mixture Model, Graph-based Image Segmentation, Foreground/Background Estimation.

1 Introduction

Nowadays, online clothing shopping becomes an attractive and convenient shopping way for millions of web users. Especially, with the emergence of social image sharing websites, such as *Pinterest*, it accelerates the progress of social and personalized e-commerce. There exist billions of diverse and beautiful clothes available on e-commerce websites, such as *Amazon*, *eBay*, and *Alibaba*. In order to attract the eyes of customers and demonstrate the actual appearance of clothes, the clothes are usually dressed by fashion models in real world and taken pictures with natural outdoor background. Therefore, a large portion of the apparel images in e-commerce websites commonly contain cluttered and complex backgrounds, which makes visual clothing search a challenging task.

Clothing segmentation and extraction, is an active research topic in computer vision and multimedia area. Its purpose is to identify and extract the clothing itself after removing the background and unrelated information. Existing clothing segmentation methods suffer from variations in colors and styles, different lighting conditions, geometric deformations, viewpoint changes, clustered backgrounds, and occlusions generated by poses or other objects. These variations are the major factors complicating the matters for clothing extraction.

In this paper, we proposed an automatic clothing extraction algorithm by combining efficient graph-based image segmentation and foreground/background estimation. It mainly consists of two phases: a coarse clothing region localization and a fine clothing extraction. An apparel image is first segmented into multiple regions using a graph-based image segmentation approach. Skin and face regions are detected to guide the clothing region localization and assist the foreground/background model estimation. Based on the human proportion, inner and outer bound regions are roughly identified, indicating the potential clothing region and background region, respectively. Gaussian mixture model is adopted to build the foreground (clothing) and background models. By taking into account the spatial relationship among pixels, the generated GMM models are refined based on the components after efficient graph-based segmentation to achieve better segmentation performance. Experiments on two datasets crawled from e-commerce website *Taobao* and Pinterest like social sharing website *Mogujie* respectively demonstrate the proposed approach improves the segmentation results. It achieves competitive performance with the classic interactive segmentation approach GrabCut, from which users designate the desire region by dragging a rectangle around the object.

The rest of paper is organized as follows. Section 2 gives a brief overview of the related work. Section 3 elaborates the proposed clothing extraction algorithm. Section 4 presents the experiments. Finally, we summarize this paper with a conclusion.

2 Related Work

2.1 Product Image Search

In industry, shopping comparison website *Like.com*, is the first product image search engine to bring visual search for shopping, which builds an automated matching system for products, such as jewelries, handbags, shoes, and watches. It exploits computer vision and machine learning techniques to find similar-looking (similar colors, shapes, and patterns) products. In China, *Taotaosou* [13] under *Alibaba*, provides similar functions for visual product search. In academic research, *iLike* [3] explores vertical search by integrating textual and visual features to improve search performance, particularly targeting for product search of apparels and accessories. *iSearch* [10] combines global and local matching of local features to find similar product images in an interactive manner. A clothes search in consumer photos is presented in [15] by color matching and attribute learning, which leverages the low-level features (colors) and high-level features (attributes) of clothes. A Smart Mirror system [1] is proposed to recognize clothing styles and supports real-time fashion recommendation. However, the above-mentioned works mainly consider the images with clean background. The situation for product images with clustered background is not considered. To handle the discrepancies between online shopping images and daily photos, a two-step cross-scenario clothing retrieval is proposed via parts alignment and auxiliary set [11].

2.2 Clothing Segmentation

Image segmentation is widely used in many image related applications, such as content-based image retrieval, image annotation, and object recognition. In the past few decades, numerous image segmentation approaches have been proposed, including minimum spanning tree, min-cut, normalized cut, mean shift, and so on. Recently, a number of researches have been conducted on clothing image segmentation. Clothing modeling and recognition adopts an And-Or graph representation to produce a large set of composite graphical templates accounting for the wide variability of cloth configurations [2]. Without any pre-defined clothing model, a clothing segmentation method using foreground and background estimation is proposed [6]. A torso area is first detected based on dominant colors determination and then the background area is determined based on the Constrained Delaunay Triangulation (CDF). Using these two areas, the foreground and background estimation is obtained to accomplish the clothing segmentation task. However, in our work, we simply use the human proportion other than CDF to determine the foreground and background areas, which is more efficient. Given multiple images of the same person wearing the same clothing, the clothing co-segmentation [5] provides a significant improvement in recognition accuracy, by analyzing the mutual information between pixel locations near the face and the identity of the person to learn a global clothing mask. A multi-person clothing segmentation algorithm [14] is proposed for highly occluded images, which combines blocking models to address the person-wise occlusions.

3 Clothing Extraction

3.1 Framework

The presence of natural backgrounds and fashion models could significantly influence the performance of clothing image search. In order to identify the clothes in images and remove the impact of backgrounds and models, we proposed an automatic clothing extraction algorithm for clothing image database. The framework is illustrated in Fig. 1. It mainly consists of two phases: a coarse clothing region localization and a fine clothing extraction. To reduce the effect of noises in images, a Gaussian filter, as a preprocessing step, is first deployed to smooth the images. As skin and face are useful priori information to help locate the clothes, face and skin detection are adopted to detect the face and skin regions. According to the face region and human body proportions (face, torso, and so on), a coarse inner region and an outer region are identified, from which the potential clothing region and background region are roughly located. A fine-granularity clothing extraction is then undergone to accurately identify the clothes. To model the statistical distribution of image pixels, Gaussian Mixture Model (GMM) is adopted to build the foreground (clothing) and background models. At the same time, efficient graph-based image segmentation [4] is applied to segment the same image into multiple components, which act as an auxiliary resource. By taking into account the neighborhood and spatial relationship among pixels, the generated GMM models are refined to achieve better segmentation performance. Finally, the clothes are extracted from images, which can be used for visual clothing search to improve the performance.

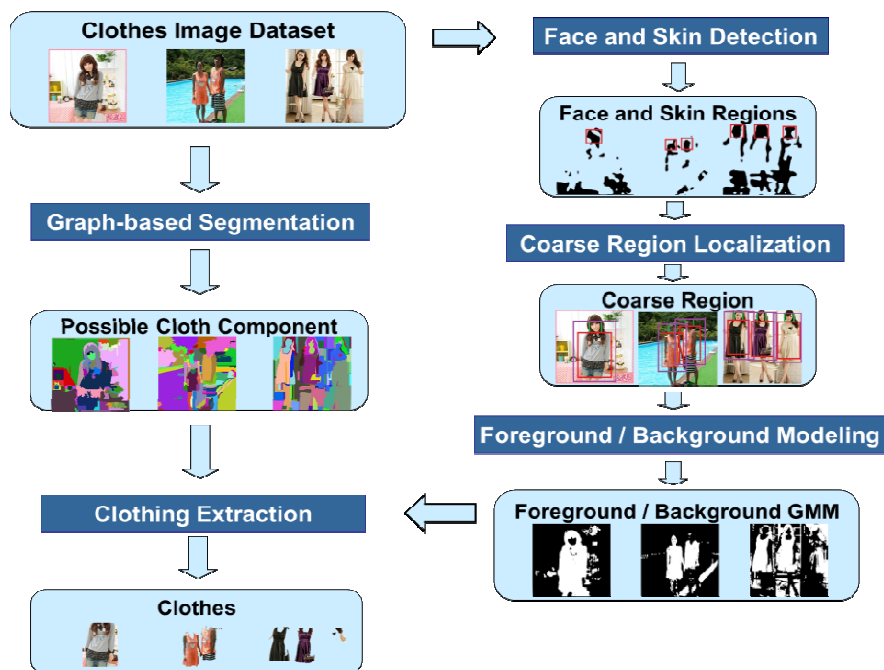


Fig. 1. Framework of the proposed clothing extraction

3.2 Skin and Face Detection

Our clothing extraction is guided by the detected faces. An Adaboost based face detector [7] is used in our work to locate the faces in different images, which has been widely used in many applications. It can accurately detect faces in real-life images with kinds of poses changes.

Skin pixels belong to the background, but sometimes some of them appear in the inner region and lie out of the determined background pixels. The wrong-classified skin pixels can affect the correct color distribution of both foreground and background, leading to poor segmentation. Skin pixels should be removed from the foreground seeds and be added into the background seeds to solve the above problem.

For skin detection, since Single Gaussian Model is sensitive to red-like pixels while Elliptical Boundary Model is sensitive to skin-like pixels [8], we combine Single Gaussian Model and Elliptical Boundary Model to obtain the skin area. A single Gaussian probability distribution using YCbCr color space is adopted to depict the skin distribution. Skin-color distribution is modeled through Gaussian joint probability distribution function. The parameters are estimated over all the color samples from the training data using Maximal Likelihood Estimation (MLE). The overlap of the results from Single Gaussian Model and Elliptical Boundary Model is treated as the final skin regions. The detected skin regions are illustrated in Fig. 2.



Fig. 2. The original images and the detected skin regions

3.3 Coarse Clothing Region Localization

As the clothing region is connected to the head, we exploit the detected face to guide the coarse clothing region localization. Though there are subtle differences between individuals, human proportions fit within a fairly standard range. In figure drawing, the basic unit of measurement is the “head”, which is reasonably standard and has long been used to establish the proportions of the human figure. According to the study, an average person is generally 7-and-a-half heads tall (including the head) [16], which is shown in Fig 3(a).

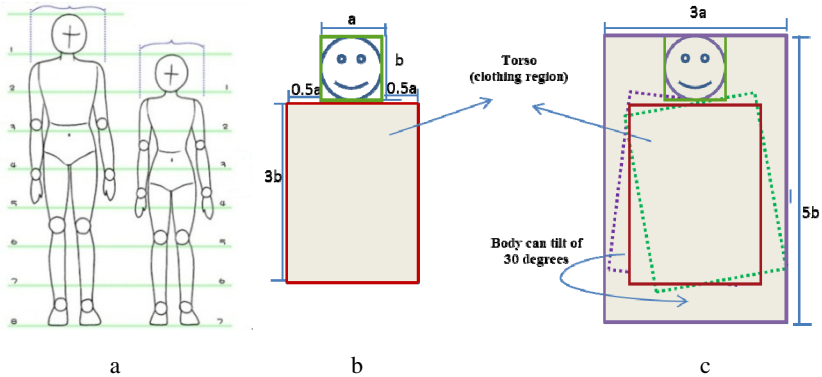


Fig. 3. Human proportions (a) and the inner bound (b) and outer bound (c)

A coarse clothing region can be firstly outlined based on the human proportions and clothing properties. Two rectangle regions called *inner bound* and *outer bound* are identified. The pixels in inner bound have high probability of belonging to the clothing, while the pixels outside the outer bound indicate the background. Assume that the width and height of the detected face are a and b , respectively, the region

positioned right below the face with the width and length ratio as 2a:3b is treated as the inner bound. The region including the face with the width and length ratio as 3a:5b is treated as the outer bound, which contains the face region and the inner bound region. The inner and outer bound are illustrated as red and purple rectangles in Fig. 3(b) and (c), respectively. Some examples with detected face, inner and outer bounds are shown in Fig. 4. These coarsely detected inner and outer bound will be exploited to construct the foreground and background models.

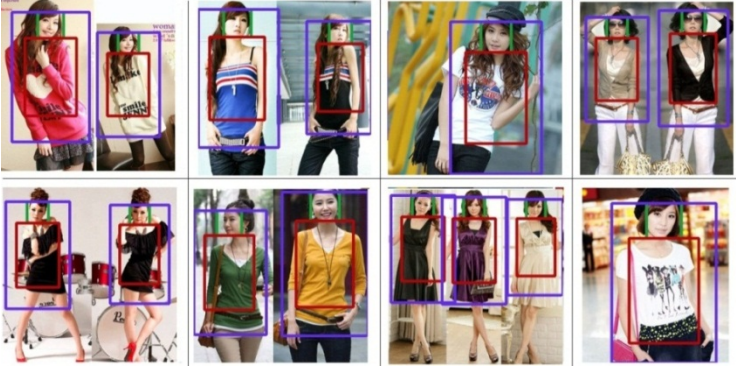


Fig. 4. The detected face, inner bound and outer bound regions

3.4 Clothing and Background Modeling

With the inner and outer bounds, the foreground (clothing) seeds are estimated from the inner region exclude the skin regions based on main colors determination, and the background (non-clothing) seeds are found based on the outer region plus skin regions. As foreground and background seeds contain several main colors, Gaussian Mixture Model (GMM) is employed to interpret color distributions of such mixture data.

Two GMMs are used to model the image color distributions of the clothing and background, respectively. In this work, the RGB color space is deployed.

$$p(x|clothes) = \sum_{i=1}^{K_c} \pi_i^c \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i^c|^{\frac{d}{2}}} \exp \left\{ -\frac{1}{2} x - \mu_i^{cT} \sum_i^{c-1} (x - \mu_i^c) \right\} \quad (1)$$

$$p(x|background) = \sum_{i=1}^{K_b} \pi_i^b \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma_i^b|^{\frac{d}{2}}} \exp \left\{ -\frac{1}{2} x - \mu_i^{bT} \sum_i^{b-1} (x - \mu_i^b) \right\} \quad (2)$$

where x is a 3D vector standing for the RGB value of pixel x , μ_i^c and Σ_i^c are the mean value and covariance matrix of the i th Gaussian of the clothing GMM, μ_i^b and Σ_i^b are the mean value and covariance matrix of the i th Gaussian of the background

GMM. π_i^c and π_i^b are weighting factors of i th Gaussian of clothing and background respectively. All these parameters are determined by EM algorithm. K_c and K_b are the number of Gaussian distributions. In our experiments, they are set as 4.

GMM considers the statistical information which means pixels with similar color have the similar probability belongs to the clothing or background, but it ignores the spatial information which means pixels near each other should have similar probability. In addition, as GMM makes good use of the pixels' color properties, it is sensitive to illumination variations and clustered colors. To alleviate this problem, GMM-based color distribution integrates the efficient graph-based image segmentation [4] to improve the segmentation performance, which combines both the color properties and region properties.



Fig. 5. Components after efficient graph-based image segmentation

To get space information, we consider the results of efficient graph-based segmentation [4] which cuts an image into several components. The detected components after efficient graph-based segmentation are shown in Fig. 5. For each component C_j after image segmentation, we calculate its foreground and background probabilities. The foreground probability $p(C_j|clothes)$ and background probability $p(C_j|background)$ are defined as the mean foreground probabilities and background probabilities of all pixels in the component, respectively, which are defined as follows:

$$p(C_j|clothes) = \frac{1}{M} \sum_{i=1}^M p(x_i|clothes) \tag{3}$$

$$p(C_j|background) = \frac{1}{M} \sum_{i=1}^M p(x_i|background) \tag{4}$$

where x_i is the i th pixel belongs to C_j and M is the total number of pixels in C_j .

The refined models $p(x_i^j|clothes)$ and $p(x_i^j|background)$ are determined by the combination of the original probability and the component probability. They consider the statistical information and spatial information, which are defined as:

$$p(x_i^j|clothes) = p(x_i|clothes) + p(C_j|clothes) \tag{5}$$

$$p(x_i^j|background) = p(x_i|background) + p(C_j|background) \tag{6}$$

The pixels are treated as the clothing pixel, if these pixels are within the outer bound region whose $p(x_i^j|clothes) > p(x_i^j|background)$.

4 Experiments

There is no public product image dataset and corresponding ground truth available for evaluating the performance of clothing extraction. To evaluate the performance, we crawled product images from *Taobao*, the biggest e-commerce website in Asia. Totally, there are 1,356,901 images. These images are mainly from two categories: clothes and handbags. Since manually labeling the ground truth of clothing extraction on a dataset with millions of images is time-consuming, it is infeasible to evaluate on the whole dataset. We use two datasets: DS_TB, and DS_MGJ to evaluate the performance of the proposed solution. DS_TB consists of 1000 images with faces randomly selected from the above-mentioned dataset as the evaluation dataset. In addition, we crawled another 1000 clothing images from a Pinterest-like website in China, *Mogujie* (www.mogujie.com), which are mainly captured from outdoors, as the DS_MGJ.

Due to without the ground truth of accurate pixel-level clothing segmentation, it is impossible to evaluate the performance in an objective way. In this work, we use subjective evaluation for the performance of clothing extraction. Based on the clothing extraction results of different algorithms, five assessors were requested to evaluate the quality of clothing extraction by giving a score between 0 and 5 to the image, indicating the accuracy of the extracted clothing comparing to the perfect extraction. A higher score means a better segmentation performance. Score 5 refers to perfect clothing extraction, while 0 indicates that none of the extracted part belongs to the clothing. We use *average accuracy score* as the performance metric, which is defined as the sum of the scores for all images to the total number of images. In our work, N is 1000.

$$\text{aas} = \sum_{i=1}^N \text{Score}_i / N \quad (7)$$

To compare the performance, we compare the proposed solution with the Principal Object Detection (POD) [17], the simplified GMM based approach [6] and the interactive Grabcut [12]. Grabcut is an interactive image segmentation solution with human interaction by dragging a rectangle region in the query image to guide the object identification. Although the user interaction scheme is impractical for large scale object extraction, we evaluate the performance of the automatic solution compared to the interactive way. The principal object detection is induced from the efficient graph-based image segmentation. Based on the intuition that the object should be in the middle of the image and the size should not be small, the component in the middle and with large region will be treated as the clothing object.

Fig. 6 demonstrates the average accuracy score of different approaches in datasets DS_TB and DS_MGJ. Overall, the proposed approach achieves the highest score compared with POD and GMM based approach in both datasets. In addition, without user interaction, the proposed solution has competitive performance as the interactive approach GrabCut. It means that our method can be applicable for large scale backend image datasets. The POD performs poor when facing images with complex backgrounds. Its performance is affected by the graph-based image segmentation. It

might select the wrong region as the principal object. It should be noted that the overall segmentation results in DS_MGJ are poorer than the ones in DS_TB. Most images in dataset DS_MGJ are captured outdoors with cluttered background, while parts of the images in DS_TB have relatively simple backgrounds. It makes the images in DS_MGJ are more challenging, which significantly affects the extraction performance. Fig. 7 shows the extracted results with different approaches. Generally, the extracted clothes using GrabCut and our method are more comprehensive and meaningful. Additionally, our algorithm is very efficient. On average, the clothing extraction process takes about 4 seconds per image on an Intel Core i5 3.1GHz processor with 4GBs of RAM.

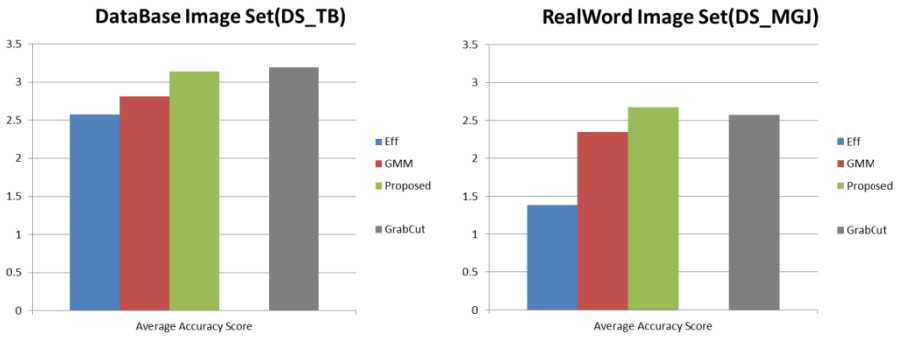


Fig. 6. Average accuracy score of different approaches in two datasets

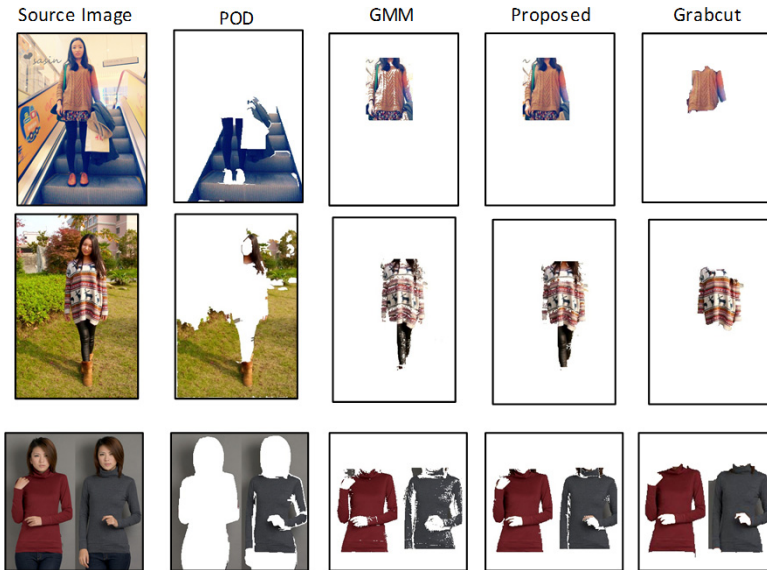


Fig. 7. Performance comparison with different approaches

5 Conclusion

In this paper, we explore the clothing extraction algorithm with two steps: coarse clothing region localization and fine clothing extraction, which automatically localize the clothing region and estimate the foreground/background models to extract the clothing. Experiments on two datasets demonstrate the effectiveness of the proposed approach. In our future work, we will exploit the spatial symmetric property and texture consistency of clothes to further improve the segmentation accuracy. In addition, we will explore the clothing co-segmentation when there exist multiple images with similar clothes. Our ultimate goal is to propose unsupervised image segmentation algorithms which can efficiently and accurately extract clothing from images with cluttered background and fashion model.

Acknowledgements. The work described in this paper was supported by the National Natural Science Foundation of China (No. 61071184, 60972111, 61036008), Research Funds for the Doctoral Program of Higher Education of China (No. 20100184120009, 20120184110001), Program for Sichuan Provincial Science Fund for Distinguished Young Scholars (No. 2012JQ0029), the Fundamental Research Funds for the Central Universities (Project no. SWJTU09CX032, SWJTU10CX08, SWJTU11ZT08), and Open Project Program of the National Laboratory of Pattern Recognition (NLPR).

References

1. Chao, X., Huiskes, M.J., Gritti, T., Ciuhu, C.: A Framework for Robust Feature Selection for Real-time Fashion Style Recommendation. In: ICME, pp. 35–41 (2009)
2. Chen, H., Xu, Z., Liu, Z., Zhu, S.: Composite Templates for Cloth Modeling and Sketching. In: CVPR (2006)
3. Chen, Y., Yu, N., Luo, B., Chen, X.-W.: iLike: Integrating Visual and Textual Features for Vertical Search. In: ACM MM, pp. 221–230 (2010)
4. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient Graph-based Image Segmentation. *IJCV* 59(2), 167–181 (2004)
5. Gallagher, A.C., Chen, T.: Clothing Cosegmentation for Recognizing People. In: CVPR (2008)
6. He, Z., Yan, H., Lin, X.: Clothing Segmentation using Foreground and Background Estimation based on the Constrained Delaunay Triangulation. *Pattern Recognition* 41, 1581–1592 (2008)
7. Jones, M.J., Viola, P.: Fast Multi-view Face Detection. In: CVPR (2003)
8. Kakumanu, P., Makrogiannis, S., Bourbakis, N.: A Survey of Skin-color Modeling and Detection Methods. *Pattern Recognition* 40, 1106–1122 (2007)
9. Lee, J.Y., Yoo, S.I.: An Elliptical Boundary Model for Skin Color Detection. *Science* (2002)
10. Li, H., Wang, X., Tang, J.-H., Yi, L., Xiao, L.: iSearch: Towards Precise Retrieval of Item Image. In: ACM ICIMCS, pp. 5–8 (2011)
11. Liu, S., Song, Z., Liu, G., Xu, C.-S., Lu, H., Yan, S.-C.: Street-to-Shop: Cross-Scenario Clothing Retrieval via Parts Alignment and Auxiliary Set. In: CVPR (2012)

12. Rother, C., Kolmogorov, V., Blake, A.: Grabcut - Interactive Foreground Extraction using Iterated Graph Cuts. *ACM SIGGRAPH* 23, 309–314 (2004)
13. Taotaosou, <http://www.taotaosou.com>
14. Wang, N., Ai, H.: Who Blocks Who: Simultaneous Clothing Segmentation for Grouping Images. In: *ICCV*, pp. 1535–1542 (2011)
15. Wang, X., Zhang, T.: Clothes Search in Consumer Photos via Color Matching and Attribute Learning. In: *ACM MM*, pp. 1353–1356 (2011)
16. Wikipedia, http://en.wikipedia.org/wiki/Body_proportions
17. Wu, X., Liang, L.-L., Wang, W.-J., Peng, Q.: Principal Object Detection towards Product Image Search. In: *ICALIP*, pp. 866–871 (2012)

Object Categorization Using Local Feature Context

Tao Sun¹, Chunjie Zhang², Jing Liu¹, and Hanqing Lu¹

¹ National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, P.O. Box 2728, Beijing, China

² School of Computer and Control, University of Chinese Academy of Sciences
{tsun, jliu, luhq}@nlpr.ia.ac.cn, cjzhang@jdl.ac.cn

Abstract. Recently, the use of context has been proven very effective for object categorization. However, most of the researchers only used context information at the visual word level without considering the context information of local features. To tackle this problem, in this paper, we propose a novel object categorization method by considering the local feature context. Given a position in an image, to represent this position's visual information, we use the local feature on this position as well as other local features based on their distances and angles to this position. The use of local feature context is more discriminative and is also invariant to rotation and scale change. The local feature context can then be combined with the state-of-the-art methods for object categorization. Experimental results on the UIUC-Sports dataset and the Caltech-101 dataset demonstrate the effectiveness of the proposed method.

Keywords: bag of visual words, local feature context, sift, object categorization.

1 Introduction

Currently, the state-of-the-art methods for object categorization are based on the information collection of local image features. A codebook is typically generated by a clustering method, such as k -means clustering. Each local feature is then quantized by nearest neighbor assignment [1]. Recent experimental results show that the object categorization performance can be improved by soft assignment such as kernel codebook [2] and sparse coding [3]. The bag-of-visual words (BoW) is used which represents an image based on the occurrences of visual words.

However, the BoW model disregards the spatial information and correlations of local features. To alleviate this problem, a simple but very efficient method called spatial pyramid matching (SPM) is proposed by Lazebnik *et al.* [4]. The SPM partitions an image into increasingly finer spatial sub-regions and uses the occurrence of visual words within this sub-region for image representation. Typically, $2^l \times 2^l$ with $l = 0, 1, 2$ is used. On the other hand, the use of context has become popular in recent years [5-9]. Wu *et al.* [5] bundled features together and applied it to web image search to improve the performance. Ni *et al.* [6] proposed contextualizing histogram to incorporate spatial contextual information into histogram based image

representation to boost the visual classification performance. Yao *et al.* [7] used mutual context to model objects and human poses for action classification. Lee and Grauman [8] constructed object-graphs to automatically discover object categories by modeling object relationships with graphs. Belongie *et al.* [9] proposed a semi local shape descriptor, called Shape Context. The Shape Context represents a binary shape as a discrete set of points sampled from its contour. These points are then mapped into a log-polar coordinate system centered at a reference point. Each bin of the log-polar space is determined by the distance and angle intervals. Although proven effective, most researchers only focused on the context modeling at the visual word level. However, the contextual information at the local feature level is seldom explored. Due to quantization loss during visual word assignment process, it would be more effective to model the contextual relationship of local features directly.

In this paper, we propose a novel object categorization method by modeling the local feature context. In order to represent the visual information of a given position in an image, we use the local feature extracted on this position as well as the other local features depending on their distances and angles to this position. This representation is more discriminative than only using local feature on this position; it is also invariant to rotation and scale change. The proposed local feature context can then be combined with the state-of-the-art methods for object categorization. Experimental results on the UIUC-Sports dataset and the Caltech-101 dataset demonstrate the effectiveness of the proposed method. Figure 1 shows the flowchart of the proposed method.

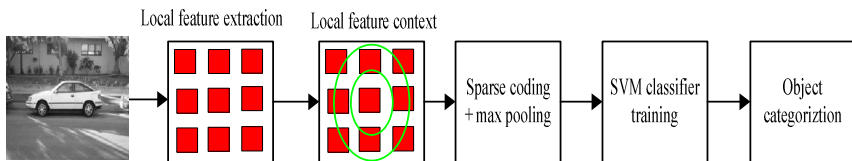


Fig. 1. Flowchart of the proposed object categorization using local feature context method

The rest of the paper is organized as follows. In Section 2, we give the details of how to use the local features context information for object categorization. In Section 3, we show the experimental results on the UIUC-Sports dataset and the Caltech-101 dataset. Finally, we conclude in Section 4.

2 Local Feature Context for Object Categorization

This section gives the details of the proposed object categorization using local feature context method. For a given position in each image, besides using the local features on this position, we also use other local features based on their distances and angles to this position. Sparse coding along with max pooling is then used to encode the local features and get the histogram representation of images. We train SVM classifier with multiple kernel learning (MKL) technique [10] to predict the categories of images.

2.1 Local Feature Context

In a given image X , suppose we have a set of N local features $X = [x_1, \dots, x_N] \in \mathbb{R}^{D \times N}$ where $x_i \in \mathbb{R}^{D \times 1}$, $i = 1, \dots, N$. The locations of these local features are $Z = [z_1, \dots, z_N]$ where $z_i \in \mathbb{R}^{2 \times 1}$, $i = 1, \dots, N$. Let p be a location of a reference point in image X , the area around p is divided into sub-regions $sub_region_r^p$ in the log polar coordinate system with $r = 1, \dots, R$, as shown in Figure 2 (with $R = 1$). R is the number of sub-regions. The local feature context (LFC) of location p is then defined as:

$$LFC(p, r, l) = f(\|z - p\| < l) \tag{1}$$

$$\forall z \in sub_region_r^p$$

where l is the distance between location z and p . f is a pooling function which extracts the visual information in the $sub_region_r^p$. The pooling function of f can be max, sum, mean or concatenation. We choose to use these four types of pooling methods in our experiments and combine their discriminative power using the multiple kernel learning technique. Usually, we have a set of reference points for each image; hence, the local feature context of an image X is defined as:

$$LFC(X) = (LFC(p, r, l))_{p, r, d \in \mathbb{R}^{p \times R \times L}} \tag{2}$$



Fig. 2. Reference point selection method used in this paper

where P is the number of reference points and L is the number of distances. We choose the reference points of P in a spatial pyramid matching way similar with [4]. For each sub-region of SPM, we choose the center of this sub-region as the reference point for this sub-region. Figure 2 shows the reference point selection using $2^l \times 2^l$ partition of images with $l = 0, 1, 2$.

2.2 Object Categorization by Local Feature Context

After the local feature context extraction, we can encode this information per pooling type for object categorization. We choose the sparse coding plus max pooling method proposed by Yang *et al.* [3] for its good performance over the traditional k -means clustering method. Formally, let $A = [a_1, \dots, a_T] \in \mathbb{R}^{Q \times T}$ be the set of Q -dimensional local context feature with the number of local context feature is T . The sparse coding tries to learn the codebook $B = [b_1, \dots, b_M] \in \mathbb{R}^{Q \times M}$ with M visual words as well as the coding parameters C by solving the optimization problem as:

$$[B, C] = \arg \min_{B, C} \sum_{t=1}^T \|a_t - B \times c_t\|^2 + \lambda \|c_t\|, \quad (3)$$

where $C = [c_1, \dots, c_T]$, λ is the parameter which controls the sparsity of C . Max pooling is then use to represented images. We learn the codebook and coding parameters for every pooling method in (1) and represent an image by a set of histograms. We then use the MKL technique to make use of this representation. The MKL tries to find a linear combination of different kernels (K_1, \dots, K_j) such that the resulting kernel $K = \sum \alpha_j K_j$ is “optimal” in some sense, where $\alpha = [\alpha_1, \dots, \alpha_j]$ is the combination parameters. In this paper, we choose the χ^2 kernel for its good performance for object categorization. Given two histograms $h_i, h_j \in \mathbb{R}^{Y \times 1}$, The χ^2 kernel is defined as:

$$k(h_i, h_j) = \sum_{y=1}^Y \frac{(h_{i,y} - h_{j,y})^2}{h_{i,y} + h_{j,y}} \quad (4)$$

After the classifier training, we can predict the object categories using the learned kernel K .

3 Experiments

We evaluate the proposed local feature context for object categorization method on two public datasets: the UIUC-Sports dataset from Li and Fei-Fei [11] and the Caltech-101 dataset from [12]. The codebook size is set to 1,024 for the two datasets,

as in [3]. We use the same setup as in [2, 3, 4] to extract local features because this setup has been proven effective on these datasets. We densely extract SIFT [13] descriptors on 16×16 pixels with an overlap of 6 pixels. We process all images in gray scale. We follow Lazebnik *et al.* [4] and use the first 3 layers for spatial pyramid matching with the same weight for each layer. The one-versus-all rule is used for multi-class classification and a SVM classifier is learned to separate each class of images from the rest images. The test images are assigned the label of classifiers with the highest responses. We use the average of per-class classification rates for quantitative performance comparison.

3.1 UIUC-Sports Dataset

The UIUC-Sports dataset has eight categories of 1,792 images with the eight categories as: badminton, bocce croquet, polo, rock climbing, rowing, sailing and snow boarding. The number of images per categories ranges from 137 to 250. Figure 3 shows some example images of the UIUC-Sports dataset. We follow the same experimental setup as in [11] and randomly choose 70 images per class for training and use the rest images for testing. This process is repeated for five times to get reliable results.



Fig. 3. Example images of the UIUC-Sports dataset

Table 1 gives the performance comparison for the proposed method and methods in [3, 11, 14] on the UIUC Sports dataset. Yang *et al.* [3] used the sparse coding along with max pooling method to extract histogram representation of images and used the spatial pyramid matching to consider the spatial information of local features. Li and Fei-Fei [11] tried to model the scene relationship using an integrative model. Wu and Rehg [14] generated the codebook using histogram intersection kernel and used a one-class SVM formulation to create more effective visual words. We can see from Table 1 that the proposed object categorization using local feature context method outperforms the ScSPM [3], TIM[11] and HIK+OCSVM [14]. This demonstrates the effectiveness of the proposed method. Our method considers the spatial information and correlations among local features, hence is more discriminative than using local features along. Besides, our method can be combined with other state-of-the-art methods (*e.g.* sparse coding) which can further improve the object categorization performance.

3.2 Caltech-101 Dataset

The Caltech-101 dataset contains 102 classes with high intra-class appearance shape variability. We exclude the background class and only use the 101 object classes for object categorization performance evaluation. The number of images per class varies from 31 to 800 images with most of the images of medium resolution. We follow the same experimental setup as did in [2, 12] for fair comparison. We randomly choose 15 and 30 images per class for classifier learning and up to 30 images per class for testing. This process is repeated for ten times.

Table 1. Performance comparison on the UIUC-Sports dataset. ScSPM: Sparse coding along with spatial pyramid matching; TIM: the integrative model; HIK+OCSVM: one-class SVM with histogram intersection kernel.

Methods	Performance
ScSPM[3]	82.74 ± 1.46
TIM[11]	73.40
HIK+OCSVM[14]	83.54 ± 1.13
LFC	84.39 ± 0.95

Table 2 gives the performance comparison for the proposed method and methods in [2, 3, 4, 15, 16] on the Caltech-101 dataset. As shown, the proposed LFC method achieves the state-of-the-art performance and outperforms ScSPM by 1 percent for 15 training images and LLC by 1.8 percent for 30 training images. This also demonstrates the effectiveness of the proposed method.

Table 2. Performance comparison on the Caltech-101 dataset. KCSPM: kernel codebook with spatial pyramid matching; SPM: spatial pyramid matching; ScSPM: sparse coding along with spatial pyramid matching; SVM-KNN: A hybrid nearest neighbor based and SVM based method; LLC: locality-constrained linear coding.

Methods	15 training	30 training
KCSPM[2]	-	64.14 ± 1.18
SPM[3]	56.40	64.40 ± 0.80
ScSPM[4]	67.00 ± 0.45	73.20 ± 0.54
SVM-KNN[15]	59.10 ± 0.60	66.20 ± 0.50
LLC[16]	65.43	73.44
LFC	68.05 ± 0.84	75.29 ± 0.92

4 Conclusion

This paper proposes a novel object categorization method by using local feature context. To represent a given image point, we use the local feature on this position as well as other local features based on their distances and angles to this position. The proposed local feature context is more discriminative and is also invariant to rotation and scale change. It can then be combined with the state-of-the-art methods for object

categorization. Experimental results on the UIUC-Sports dataset and the Caltech-101 dataset demonstrate the effectiveness of the proposed method.

Acknowledgments. This work was supported by 973 Program(2010CB327905) and National Natural Science Foundation of China(61070104, 61025011, 61272329).

References

1. Sivic, J.S., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proc. ICCV, vol. 2, pp. 1470–1477 (2003)
2. Gemert, J., Veenman, C., Smedulders, A., Geusebroek, J.: Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2010)
3. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: Proc. CVPR (2009)
4. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proc. CVPR (2006)
5. Wu, Z., Ke, Q., Sun, J.: Bundling features for large scale partial-duplicate web image search. In: Proc. CVPR (2009)
6. Ni, B., Yan, S., Kassim, A.: Contextualizing histogram. In: Proc. CVPR (2009)
7. Yao, B., Khosla, A., Fei-Fei, L.: Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses. In: Proc. ICML (2009)
8. Lee, Y., Grauman, K.: Object-graphs for context-aware category discovery. In: Proc. CVPR (2010)
9. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2002)
10. Bach, F., Lanckriet, G., Jordan, M.: Multiple kernel learning, conic duality, and the SMO algorithm. In: Proc. ICML (2004)
11. Li, L.J., Fei-Fei, L.: What, where and who? Classifying events by scene and object recognition. In: Proc. ICCV, Rio de Janeiro, Brazil, October 14-20 (2007)
12. 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. ICCV Workshop on Generative-Model Based Vision (2004)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. In: Proceeding of ECCV Workshop on Statistical Learning in Computer Vision, vol. 60(2), pp. 91–110 (2004)
14. Wu, J., Rehg, J.M.: Beyond the Euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In: Proc. ICCV, Kyoto, Japan (2009)
15. Zhang, H., Berg, A., Maire, M., Malik, J.: SVM-KNN: Discriminative nearest neighbor classification for visual category recognition. In: Proc. CVPR (2006)
16. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: Proc. CVPR (2010)

Statistical Multiplexing of MDFEC-Coded Heterogeneous Video Streaming*

Hang Zhang, Adarsh K. Ramasubramonian, Koushik Kar, and John W. Woods

Rensselaer Polytechnic Institute, Troy NY 12180, USA
{zhangh10,ramasa}@rpi.edu, {koushik,woods}@ecse.rpi.edu

Abstract. In this paper, we propose an approach that combines statistical multiplexing and Multiple Description with Forward Error Correction (MDFEC) coding to make the optimal use of server bandwidth, taking into account the dynamically evolving content (complexity) of the different videos being streamed by the server, and path bandwidths and loss rates experienced by the users. We formally pose and analyze the complexity of the MDFEC statistical multiplexing problem, and present a dynamic programming based polynomial-time algorithm to compute the optimum solution. We also evaluate the performance of the proposed approach against those that do not use either MDFEC or statistical multiplexing, based on the experimental results obtained from real video sequences. Besides optimizing the overall distortion across all users, our approach is quite effective in providing differentiation between user groups with significantly different path bandwidths, particularly when a weighted version of our approach is used.

1 Introduction

The last decade has seen a tremendous growth in the demand of social media services, including social networking websites (such as Twitter and Facebook), photo and video sharing websites (such as Flickr and Youtube). Video streaming applications become more and more important in social media with the development of broadband networks [1]. The set of video requests and receivers, the content (coding complexity) of each video, and the source-to-receiver path characteristics, are typically time varying. This implies that equal or static allocation of the server bandwidth among the different videos, and fixed-rate coding of the individual videos, are in general not optimal. Statistical multiplexing refers to the technique of dynamically assigning compression bit-rates to the different videos so the encoder of a more complex video is allowed to borrow bandwidth from the encoder of a less complex one [2]. This provides better overall performance at the cost of extra computational complexity [3].

Multiple Description coding with *Forward Error Correction* (MDFEC), introduced in [4], is a promising technology that provides easy adaptivity and distortion-rate optimality - which are necessary or desirable requirements for

* This work was sponsored by National Science Foundation under Grant CNS-1018398.

delivering streaming video in dynamic network environments with time-varying receiver populations and path bandwidths. With MDFEC coding, video is coded as multiple *descriptions*, and different parts of the video are protected from channel losses through differentiated redundancy (FEC) provisioning. MDFEC code construction ensures that the video quality at the receivers depends only on the *number* of distinct packets (descriptions) received, and not on *which* packets (descriptions) are received.

In this work we propose and evaluate a video streaming approach that uses both MDFEC and statistical multiplexing in a network model that contains a server with video coding capability and a pool of receivers with different path/access link bandwidths and losses, each interested in receiving one of several videos being offered (streamed) by the server. The amount of server bandwidth assigned to each video is based on the complexity of the video content. Each video, which is multicast to the different users interested in the video, is coded using MDFEC on a GOP-by-GOP basis using experimentally derived distortion-rate characteristics particular to that GOP. Thus, in our solution, while statistical multiplexing is used to address the heterogeneity among the different videos, MDFEC coding of each video is used to account for the heterogeneity in receiver capabilities and path characteristics. While statistical multiplexing for layered multicasting has been considered recently in [5], there are several key differences between this prior work and ours. Our use of MDFEC implies FEC-protection of video data against losses, which is not inherently provided by layered multicasting. Moreover, we approach the statistical multiplexing problem from the perspective of minimizing the overall distortion – a reasonable measure of the Quality-of-Experience (QoE) aggregated over all users – not considered in previous work.

The rest of the paper is organized as follows. Section 2 contains the problem statement of the MDFEC statistical multiplexing problem. Section 3 analyzes the problem and its MDFEC subproblem and provides a dynamic programming based solution approach. Sections 4 and 5 discuss simulation results of the proposed approach, based on experiments with real video sequences, comparing the MDFEC statistical multiplexing approach against those that do not use either MDFEC or statistical multiplexing. Section 6 lists our conclusions as well as possibilities for future work.

2 Problem Formulation

We first provide a brief overview of MDFEC. Multiple description (MD) coding [6] involves splitting the source data into two or more descriptions in such a way that even if a subset of descriptions is received, the receiver would still be able to decode the video, albeit at a lower quality. Priority encoded transmission (PET) [7] was introduced to improve the transmission of priority ordered data, e.g. the *I*, *P*, and *B* frames of MPEG2, on lossy packet networks, by generating MD codes with the help of parity bytes. The MDFEC algorithm [4] was developed to generate descriptions that are distortion-optimal for a video source over a single

lossy link between a source and a receiver. For this purpose, one can use Reed-Solomon codes (which satisfy the Maximal Distance Separable or MDS property) of type (N, n) , for $n = 1, \dots, N$ where N is the number of descriptions that we generate per GOP. As illustrated in Figure 1, the RS encoding for each section is done vertically and the FEC bytes are arranged below the corresponding input source symbols. If the receiver obtains n descriptions, then it will be able to decode all the source data up to rate R_n (the first n sections). MDFEC video coding nicely adapts itself to changes in the available capacities and the packet loss rates. The optimization algorithm returns the rate break-points $\{R_n\}_{n=1}^N$ that would minimize the distortion seen by the receiver, when the loss statistics of the link connecting the source and the receiver are known. In our problem setup, there are K MDFEC coded videos. For video k , $k = 1, \dots, K$, therefore, the above N and R_n become N_k and R_{k,n_k} respectively.

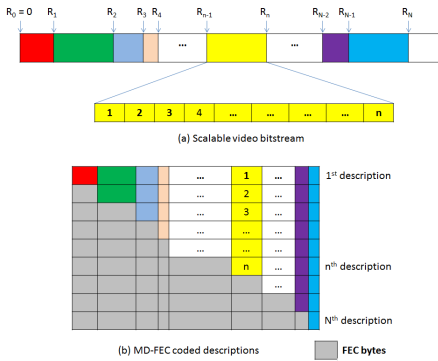


Fig. 1. MDFEC coding basics: encode a scalable video bitstream in (a) into N descriptions in (b) using Reed-Solomon (N, n) code

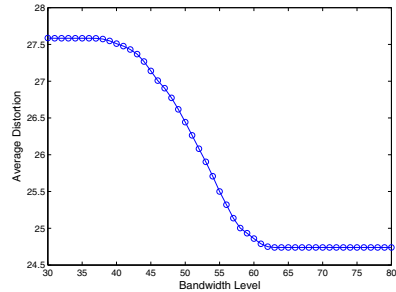


Fig. 2. The function $F_1^*(N_k)$ vs N_k for the Foreman video

In our model scalable video bitstreams are coded into a certain number of descriptions, where each description is of rate Δ . The total upload bandwidth of the server is also expressed as a multiple of Δ , i.e. $N\Delta$, where N is an integer. This server bandwidth must be allocated among K videos being streamed by the server. We assume discrete bandwidth levels, where the minimum bandwidth granularity is Δ and all receivers have path bandwidths in multiples of Δ . Let $M_k\Delta$ be the highest bandwidth of receivers interested in watching video k ($k \in \{1, \dots, K\}$), and m_{k,n_k} be the number of receivers of video k with bandwidth of $n_k\Delta$ ($n_k \in \{1, \dots, M_k\}$). Then we define the density distribution of receivers as $\rho_{k,n_k} = \frac{m_{k,n_k}}{\sum_{j=1}^K \sum_{i=1}^{M_j} m_{j,i}}$, so ρ_{k,n_k} satisfies $\sum_{k=1}^K \sum_{n_k=1}^{M_k} \rho_{k,n_k} = 1$.

Suppose video k is assigned $N_k\Delta$ amount of the server bandwidth, where N_k is an integer. Clearly we have $\sum_{k=1}^K N_k \leq N$. In our model, we consider both (i) apparent losses that happen only due to bandwidth limitations of the paths from the source to the receivers, and (ii) additional random losses due to faulty

links (as in wireless networks), buffer overflows, etc. A receiver for video k with path bandwidth of $n_k\Delta$ would suffer an apparent loss of $(N_k - n_k)$ (if $n_k < N_k$) packets when the server assigns N_k units of bandwidth to video k . In practice, the rates of additional random losses are typically small (say 10% or less), and the probability that receiver of video k with bandwidth $n_k\Delta$ can receive i packets is denoted by $P_{k,n_k,i}$.

It is possible that N_k is less than M_k , in which case the receivers that are able to receive more than N_k packets would receive N_k packets as well. Then by MDFEC property they would get same source rate too, so we have $R_{k,N_k} = R_{k,N_k+1} = \dots = R_{k,M_k}$. If $N_k > M_k$, then the receiver with the largest path bandwidth can receive up to M_k packets, so sending M_k packets suffices. Let $D_k(R)$ be the distortion-rate function of video k . Then the density-normalized distortion of receivers with bandwidth $n_k\Delta$ for video k would be $\rho_{k,n_k} \sum_{i=1}^{n_k} P_{k,n_k,i} D_k(R_{k,i})$. Our objective is to minimize the overall average distortion D_{avg} :

$$\min_{\{N_k, R_{k,n_k}\}} D_{avg} = \sum_{k=1}^K \sum_{n_k=1}^{M_k} \rho_{k,n_k} \sum_{i=1}^{n_k} P_{k,n_k,i} D_k(R_{k,i}), \quad (1)$$

subject to $\sum_{k=1}^K N_k \Delta \leq N \Delta$. Also, for every $k \in \{1, \dots, K\}$,

$$\sum_{n_k=1}^{M_k} \beta_{k,n_k} R_{k,n_k} \leq \Delta, \quad (2)$$

$$R_{k,1} \leq R_{k,2} \leq \dots \leq R_{k,N_k} = \dots = R_{k,M_k}. \quad (3)$$

In Equation (2), $\beta_{k,n_k} = \frac{1}{n_k(n_k+1)}$ for $1 \leq n_k \leq M_k - 1$, and $\beta_{k,M_k} = \frac{1}{M_k}$ are the coefficients of the rates for video k .

3 Problem Analysis

We first analyze the structure of the optimization problem posed in Section 2. Let D_{avg}^* be the optimal solution. Then from Equation (1) we have

$$\begin{aligned} D_{avg}^* &= \min_{\{N_k, R_{k,n_k}\}} \sum_{k=1}^K \sum_{n_k=1}^{M_k} \rho_{k,n_k} \sum_{i=1}^{n_k} P_{k,n_k,i} D_k(R_{k,i}) \\ &= \min_{\{N_k, R_{k,n_k}\}} \sum_{k=1}^K \sum_{n_k=1}^{M_k} \sum_{i=1}^{n_k} \rho_{k,n_k} P_{k,n_k,i} D_k(R_{k,i}) \\ &= \min_{\{N_k, R_{k,n_k}\}} \sum_{k=1}^K \sum_{i=1}^{M_k} \sum_{n_k=i}^{M_k} \rho_{k,n_k} P_{k,n_k,i} D_k(R_{k,i}) \\ &= \min_{N_k} \sum_{k=1}^K \min_{R_{k,n_k}} \sum_{i=1}^{M_k} D_k(R_{k,i}) \sum_{n_k=i}^{M_k} \rho_{k,n_k} P_{k,n_k,i} \\ &= \min_{N_k} \sum_{k=1}^K \left[\min_{R_{k,n_k}} \sum_{i=1}^{M_k} \tilde{\rho}_{k,i} D_k(R_{k,i}) \right], \end{aligned}$$

where $\min_{R_k, n_k} \sum_{i=1}^{M_k} \tilde{\rho}_{k,i} D_k(R_{k,i})$ forms a MDFEC subproblem with constraints as described in Equations (2) and (3), and $\tilde{\rho}_{k,i} = \sum_{n_k=i}^{M_k} \rho_{k,n_k} P_{k,n_k,i}$ is the equivalent weighting factor. Let the minimal distortion of the MDFEC subproblem for video k , which is a function of N_k (the server bandwidth assigned to the video), be denoted by $F_k^*(N_k)$. Then given N_k , $F_k^*(N_k)$ can be found by an efficient $O(N)$ algorithm [4].

$F_k^*(N_k)$ in general may not be convex. As shown in Figure 2, we construct an example to demonstrate this using the third GOP of *Foreman* video sequence (video 1). In the example, N_1 increases from 30 to 80, and Δ is 10 Kbps. We assume i.i.d. binomial losses with loss rate of 10% for all the receivers. The receiver population is distributed across 100 bandwidth levels, and density distribution of receivers is:

$$\rho_{1,n_1} = \begin{cases} 0 & \text{for } n_1 = 1, \dots, 30, 41, \dots, 70 \\ \frac{1}{40} & \text{for } n_1 = 31, \dots, 40, 71, \dots, 100 \end{cases} \quad (4)$$

Since $F_k^*(N_k)$ is not convex, we can not directly use a convex programming method to solve the overall problem. Instead we present a dynamic programming based polynomial-time algorithm to compute the optimum solution.

Algorithm StatMux-MDFEC:

- (1) Initialization: Initialize two $(N + 1) \times (K + 1)$ matrices $J(\cdot, \cdot)$ and $F(\cdot, \cdot)$ as,
 - $J(0, 0) = 0, J(1, 0) = 0, \dots, J(N, 0) = 0;$
 - $J(0, 1) = \infty, \dots, J(0, K) = \infty.$
 - $F(0, 0) = 0, F(1, 0) = 0, \dots, F(N, 0) = 0;$
 - $F(0, 1) = \infty, \dots, F(0, K) = \infty;$
 - For $1 \leq n \leq N, 1 \leq k \leq K, F(n, k) = F_k^*(n).$
- (2) Iterative update: For $1 \leq n \leq N, 1 \leq k \leq K,$

$$J(n, k) = \min \begin{cases} J(n, k - 1) + F(0, k), \\ J(n - 1, k - 1) + F(1, k), \\ \dots \\ J(0, k - 1) + F(n, k). \end{cases}$$

- (3) Output minimal average distortion $J(N, K)$.

Proposition 1. *On termination of StatMux-MDFEC, $J(N, K)$ corresponds to the minimum average distortion that can be attained by any statistical multiplexing bandwidth assignment with MDFEC coding.*

In the above algorithm $J(n, k)$ represents the minimal distortion of assigning $n\Delta$ amount of server bandwidth to the first k videos. $F(n, k)$ represents the minimal distortion after solving an MDFEC problem of assigning $n\Delta$ amount of server bandwidth to video k . The iterative step (2) obtains the minimal distortion of the first k videos by comparing the distortions of the first $k - 1$ videos for different bandwidth levels ($n - i$, for $i = 0, \dots, n$) and adding to that the distortion of the k th video for the rest of the bandwidth (i levels, $i = 0, \dots, n$). We now consider

the computation time of the algorithm. Step (1) requires NK computations of a convex programming problem (MDFEC computation for a single video to determine $F_k^*(n)$) for initialization of matrix F , so it requires $O(N^2K)$ in total. Iterations in step (2) require $O(N^2K)$ calculations as well. So the total computation time is $O(N^2K)$.

In order to see the benefits from both statistical multiplexing and MDFEC, we compare the proposed approach, which we call Statistical multiplexing MD-FEC (*StatMux-MDFEC*), with the following two approaches. In order to observe the benefits from statistical multiplexing, we compare it with Half-Half MDFEC (*HH-MDFEC*), in which the total bandwidth of the server is allocated among the videos equally, and then MDFEC is applied to each video. The computation time in this case equals that of solving K MDFEC problems, which is $O(NK)$. In order to observe the benefits from MDFEC, we compare StatMux-MDFEC with Statistical Multiplexing Unirate (*StatMux-Unirate*), in which the total bandwidth of the server is still divided in a distortion-optimal way, but each video is coded at a constant rate. If the bandwidth of the receiver is higher than or equal to the rate, then it can decode the video; otherwise it can not. The StatMux-Unirate problem can be solved in polynomial time using a procedure similar to the dynamic programming algorithm described above, with computation time of $O(N^2K)$.

4 Experimental Results and Comparative Evaluation

In the experiments we use two video sequences of CIF@30(352×288) resolution: *Foreman* and *Akiyo*; therefore $K = 2$. The videos are coded into scalable bit-streams on a GOP-by-GOP basis using the enhanced MC-EZBC scalable video coder [8]. The server bandwidth varies between 400 Kbps and 1.8 Mbps with a steplength of 100 Kbps. We assume $\Delta = 10$ Kbps, so the number of descriptions the server can send varies between 40 and 180. We use the receiver density distribution in Equation (4) for both *Foreman* and *Akiyo*. Accordingly there are two clusters of receivers: the low-end receivers distributed evenly from bandwidth levels 31 to 40, and the high-end receivers distributed evenly from bandwidth levels 71 to 100.

In the following, performance is measured in terms of average peak signal-to-noise ratio (*PSNR*), which is popularly used to quantify video quality. The average PSNR measure is equivalent to the average distortion (D) measured in terms of mean square error (*MSE*), and the two are related as: $PSNR = 10 \log_{10}(255^2/D)$. We assume that all packet losses follow an i.i.d. binomial distribution with the same loss rate for all receivers.

We have obtained the results for (i) apparent losses only, where packet losses are only due to receiver path bandwidth limitation; (ii) apparent losses as well as additional random losses, with additional loss ratio of 5% and 10%. Since the performance results in these cases were similar in nature, we only show the results for the case with apparent losses as well as additional 10% random losses.

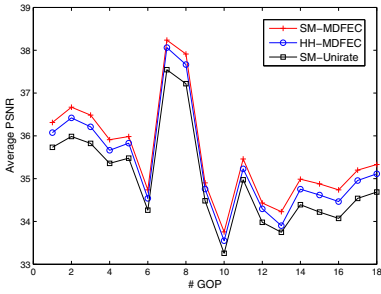


Fig. 3. The comparison of three different strategies for 10% loss rate and 800 Kbps server bandwidth

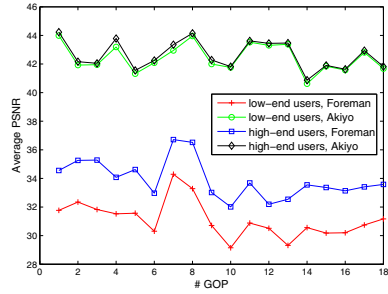


Fig. 4. Average PSNR in StatMux-MDFEC for 10% loss rate and 800 Kbps server bandwidth

4.1 Results for Different GOPs for a Fixed Server Bandwidth

Firstly we present the results for each GOP when the server bandwidth is 800 Kbps. As shown in Figure 3, the average PSNR (obtained across all users over all videos) of StatMux-MDFEC is the best, followed by HH-MDFEC, and StatMux-Unirate attains the poorest average PSNR. Besides the overall performance, we are also interested in the performance of high-end and low-end receivers of Foreman and Akiyo, as we show next.

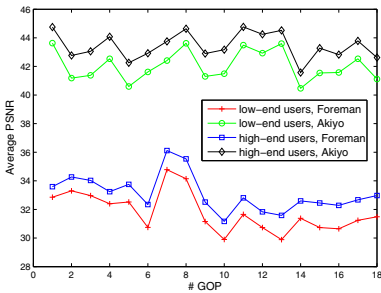


Fig. 5. Average PSNR in HH-MDFEC for 10% loss rate and 800 Kbps server bandwidth

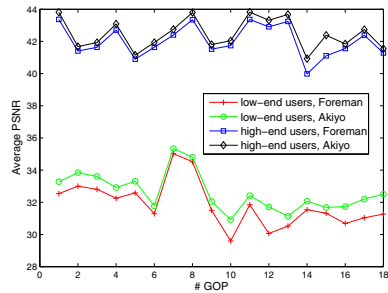


Fig. 6. Average PSNR in StatMux-Unirate for 10% loss rate and 800 Kbps server bandwidth

As shown in Figure 4, in StatMux-MDFEC, the high-end and low-end receivers of Foreman have different average PSNR, and the difference is about 2 to 3 dB. But the average PSNR of the high-end and low-end receivers of Akiyo are very close. As shown in Figure 5, in HH-MDFEC, the high-end and low-end receivers of both Foreman and Akiyo videos split into different average PSNR with the difference of 1 to 2 dB. In StatMux-Unirate, however, as shown in Figure 6, both high-end and low-end receivers attain close average PSNR for each of the

two videos. In all three approaches, Akiyo gets better performance than Foreman as it contains less motion. Since Akiyo users have already obtained much better performance (in terms of average PSNR) than Foreman, it is desirable to assign more server bandwidth to Foreman users. For StatMux-MDFEC, the server assigns 70% of its bandwidth to Foreman and 30% to Akiyo. StatMux-MDFEC also provides more differentiation between high-end and low-end receivers for the Foreman video. While the other two approaches provide more differentiation between high-end and low-end receivers for Akiyo, that hardly translates to a perceptible difference in the video quality as the PSNR for Akiyo is quite high across all users. In conclusion, StatMux-MDFEC performs better than HH-MDFEC and StatMux-Unirate in the following sense: its average PSNR is higher than the other two; it improves the performance of more complex video (Foreman), as compared to HH-MDFEC, due to statistical multiplexing effects; it provides better differentiation (compared to the other two approaches) between high-end and low-end receivers for the more complex video (Foreman) through a combination of MDFEC and statistical multiplexing effects.

4.2 Results for Different Server Bandwidths

We next present the average PSNR as calculated over all 18 GOPs. As shown in Figure 7, the average PSNR of the three approaches increases as the server bandwidth increases, and StatMux-MDFEC performs better than the other two approaches. When the server bandwidth is 400 Kbps, the average PSNR of StatMux-MDFEC is close to that of StatMux-Unirate, so MDFEC does not provide much performance benefit at this point. But we can see some effects of StatMux since Foreman is assigned more bandwidth than Akiyo is, as shown in Figure 8. This results in the significantly better performance of StatMux-MDFEC over HH-MDFEC as we see in Figure 7. As the server bandwidth increases from 400 Kbps to 600 Kbps, the bandwidth assigned to Akiyo keeps increasing as adding bandwidth to Foreman in this range does not improve its PSNR substantially, as we observe from Figure 2. When the server bandwidth is around 600 Kbps, Foreman and Akiyo get same amount of server bandwidth, so StatMux does not give

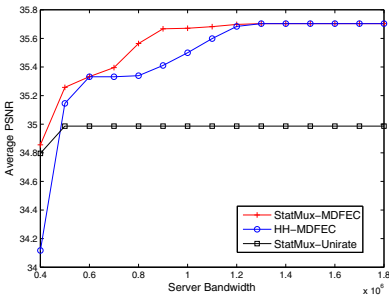


Fig. 7. The comparison of three different strategies for 10% loss rate and different server bandwidths

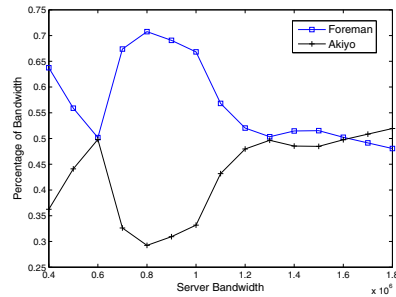


Fig. 8. The percentage of bandwidth assignment of StatMux-MDFEC for 10% loss rate and different server bandwidths

us any benefits at the point. Then as the server bandwidth increases from 600 Kbps, Foreman gets more bandwidth than Akiyo again, and StatMux-MDFEC gets better performance than both HH-MDFEC and StatMux-unirate, so both StatMux or MDFEC provides performance benefits.

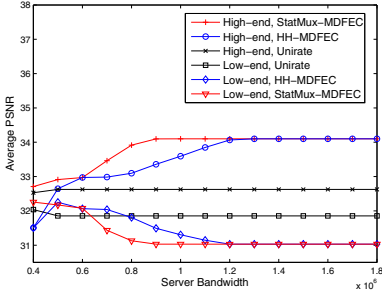


Fig. 9. Average PSNR for high-end and low-end receivers of Foreman under different strategies for 10% loss rate and different server bandwidths

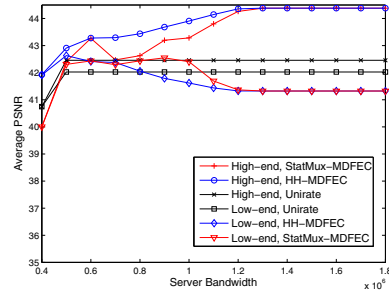


Fig. 10. Average PSNR for high-end and low-end receivers of Akiyo under different strategies for 10% loss rate and different server bandwidths

When the server bandwidth is around 800 Kbps, Foreman is assigned about 70 percent of server bandwidth, and at the same time, StatMux-MDFEC results in much better performance than StatMux-Unirate. It means that both MDFEC and StatMux work together very well at the point. As the server bandwidth increases to 1.2 Mbps and beyond, Foreman and Akiyo get similar amount of server bandwidth (about 0.6 Mbps each), and the average PSNR of StatMux-MDFEC does not improve over HH-MDFEC. This is consistent with Figure 2, where we observe that the distortion of Foreman does not improve much when the bandwidth assigned to it increases beyond 0.6 Mbps. So StatMux gives us little improvement in that range.

We would also like to study the performance of the different approaches for each video separately. As shown in Figure 9, when the server bandwidth is 400 Kbps, the average PSNR for StatMux-MDFEC and StatMux-Unirate are very close - for both high-end and low-end Foreman receivers - so MDFEC does not offer much benefit at this point. When the server bandwidth reaches 600 Kbps, StatMux-MDFEC provides the same performance for all Foreman receivers as HH-MDFEC, and they perform better than StatMux-Unirate. The performance of StatMux-Unirate does not improve as the server bandwidth increases because low-end receivers dominate the performance. But as the server bandwidth increase from 600 Kbps to 1.2 Mbps, StatMux-MDFEC provides more differentiation than HH-MDFEC does. When the server bandwidth reaches 1.2 Mbps and beyond, the curves of average PSNR of Foreman high-end and low-end receivers in HH-MDFEC reach those of StatMux-MDFEC, which also means StatMux does not give us any benefits over HH-MDFEC when the server bandwidth is sufficient. Then as shown in Figure 10, all three approaches provide high average PSNR for both high-end and low-end Akiyo receivers.

5 Statistical Multiplexing with a Weighted Distortion-Rate Function

In Section 4 we have observed that the StatMux-MDFEC approach can effectively differentiate between videos based on their content complexity, and between receivers with different bandwidth levels. In this section we propose a weighted version of StatMux-MDFEC approach to provide further differentiation between high and low bandwidth receivers.

In the unweighted case which we have considered so far, the optimization objective is Equation (1). In the weighted case that we consider next, we use the distortion of every bandwidth level as a weight factor, and then the objective becomes: $\min_{\{N_k, R_k, n_k\}} \sum_{k=1}^K \sum_{n_k=1}^{M_k} \omega_{k,n_k}(\alpha) \rho_{k,n_k} \sum_{i=1}^{n_k} P_{k,n_k,i}$, where $\omega_{k,n_k}(\alpha) = \frac{(\frac{1}{D_k(n_k \Delta)})^\alpha}{\frac{1}{M_k} \sum_{j=1}^{M_k} (\frac{1}{D_k(j \Delta)})^\alpha}$. Here α is a control parameter, and we choose to vary α between 0 and 1. When α is equal to 0, $\omega_{k,n_k}(\alpha) = 1$, so the objective becomes the standard StatMux-MDFEC objective. As α increases, the weightage provided to the $(\frac{1}{D_k(n_k \Delta)})$ term increases. Note that $D_k(n_k \Delta)$ represents the minimal distortion that receiver with bandwidth level n_k (and interested in receiving video k) can attain. Therefore, our weighting function $\omega(\cdot)$ provides more weightage to users which should have attained lower distortion in a scenario where there are no server bandwidth constraints. We present the experimental results of weighted objective when α is equal to 1. In the results shown next, the experimental parameters are similar to the unweighted case as presented earlier.

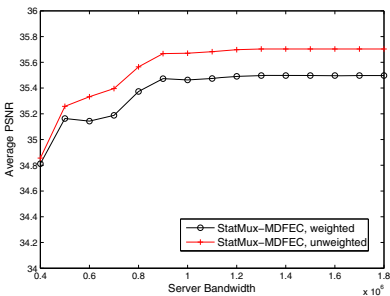


Fig. 11. The comparison of average PSNR between unweighted and weighted cases

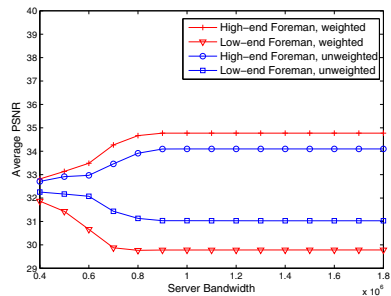


Fig. 12. Average PSNR of high-end and low-end receivers of Foreman for unweighted and weighted cases

We compare the average PSNR of weighted case with that of unweighted case, as shown in Figure 11. The average PSNR of weighted case is only 0.2 dB worse than that of unweighted case, which is reasonable since we use a new optimization objective which is not directly aimed at minimizing the PSNR but a weighted version of it. Besides, the results of bandwidth assignment in the weighted case were observed to be quite similar to those in the unweighted case shown in Figure 8.

We would also like to compare the average PSNR of high-end and low-end receivers for each video in the weighted case with that in the unweighted case. As shown in Figure 12, the high-end receivers of Foreman get about 0.7 dB advantage in average PSNR and low-end receivers of Foreman lose about 1.1 dB average. In addition, it was also observed that the average PSNRs of high-end and low-end Akiyo receivers get similar amount of differentiation to those of Foreman. Therefore the weighted approach can provide more differentiation between two clusters of receivers.

6 Conclusion and Future Work

Both StatMux and MDFEC work synergistically over a useful range of server bandwidth. The more complex video is assigned more server bandwidth than the less complex one, and at the same time receivers with greater path bandwidth get higher PSNR than receivers with lower path bandwidth. Besides, the overall average performance, which is measured by average PSNR, is also optimized on a GOP basis. The weighted StatMux-MDFEC approach can provide even more differentiation between high-end and low-end receivers without hurting the overall performance very much, i.e. less than 0.2 dB. In the future we may consider applying the proposed approach in a network environment where the packet-loss rates vary over time. We also plan to use different weight factors to better control the Quality of Experience (QoE) of the receivers.

References

1. Web could collapse as video demand soars. *Daily Telegraph* (2008)
2. Jacobs, M., Babarien, J., Tondeur, S., Van de Walle, R., Paridaens, T., Schelkens, P.: Statistical multiplexing using SVC. In: *Proc. IEEE Int. Symp. Broadband Multimedia Systems and Broadcasting*, pp. 1–6 (2008)
3. Balakrishnan, M., Cohen, R.: Global Optimization of Multiplexed Video Encoders. In: *Proceedings of ICIP*, pp. 377–380 (1997)
4. Puri, R., Ramchandran, K.: Multiple description source coding using forward error correction codes. In: *Proc. 33rd Asilomar Conf. on Signals, Systems, and Comp.*, vol. 1, pp. 342–346 (1999)
5. Jeong, J., Jung, Y.H., Choe, Y.: Statistical Multiplexing using Scalable Video Coding for Layered Multicast. In: *Broadband Multimedia Systems and Broadcasting* (2009)
6. Goyal, V.K.: Multiple description coding: compression meets the network. *IEEE Signal Processing Magazine* 18(5), 74–93 (2001)
7. Albanese, A., Blomer, J., Edmonds, J., Luby, M., Sudan, M.: Priority encoded transmission. *IEEE Trans. Inform. Theory* 42(6), 1737–1744 (2006)
8. Wu, Y., Hanke, K., Rusert, T., Woods, J.: Enhanced MC-EZBC scalable video coder. *IEEE Trans. Circuits Syst. Video Technol.* 10, 1432–1436 (2008)

Related HOG Features for Human Detection Using Cascaded Adaboost and SVM Classifiers

Hong Liu, Tao Xu, Xiangdong Wang, and Yueliang Qian

Key Laboratory of Intelligent Information Processing &&
Beijing Key Laboratory of Mobile Computing and Pervasive Device
Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190
{hliu, xutao, xdwang, ylqian}@ict.ac.cn

Abstract. Robust and fast human detection in static image is very important for real applications. Although different feature descriptors have been proposed for human detection, for HOG descriptor, how to select and combine more distinguish block-based HOGs, and how to simultaneously make use of the correlation and the local information of these selected HOGs still lack enough research and analysis. In this paper, we present a set of Related HOG (RHOG) features, including distinctive block-based HOGs (Ele-HOGs) which are selected by Adaboost and a global HOG descriptor which is concatenated by Ele-HOGs (CSele-HOG). Ele-HOG can discriminatively describe local distribution of human object while CSele-HOG contains global information. In addition, we propose a novel human detection framework of Cascaded Adaboost and SVM classifiers (CAS) based on RHOG features, which combines the advantages of Adaboost and SVM classifiers. Experimental results on INRIA dataset demonstrate the effectiveness of the proposed method.

Keywords: Machine Learning, Human Detection, Cascade, Adaboost, HOG.

1 Introduction

Human detection is an essential task in visual surveillance, image/video retrieval, and video annotation. However, detecting humans is a challenging problem for people's variable appearance, poses, clothes, illumination and complex background, especially in static image without motion information.

Machine-learning and sliding-window based human detection systems are presently the predominant methods [1, 5, 7]. In these approaches, each image is densely scanned from the top left to the bottom right with a rectangle sliding window in different scales. For each sliding window, certain features are extracted and sent to a classifier, which is trained offline on labeled training data [5]. It classifies the sliding window as human or nonhuman. For accurate human detection in real applications, the selection of feature descriptors and the classification algorithm are important factors.

Many feature descriptors have been proposed for human detection. Papageorgiou *et al.* [11] used Haar-like feature to describe different objects, such as faces, people

and cars. This feature was proved to be less effective for human detection than for face detection. Then some other feature descriptors were proposed [1, 6, 9, 12], among which Histograms of Oriented Gradients (HOG) [1] is considered as one of the most successful human descriptor. In recent years, many variants of HOG [5, 16, 17, 19, 21] have been presented to improve performance of accuracy and speed. Besides edge and gradient feature, many researchers combined different kinds of features, e.g. Duan *et al.* [10] proposed Associated Pairing Comparison Features to combine color and gradient information. Ye *et al.* [3] designed a set of multi-scale orientation features which contains coarse and fine features. Combinations of HOG and Local Binary Pattern (LBP) feature were also proposed for human detection [5, 14, 15].

Although different feature descriptors have been proposed for human detection, for original HOG descriptor, how to select and combine more distinguishable block-based HOGs, and how to simultaneously make full use of the correlation and the local information of these selected HOGs lack enough research and analysis.

Besides the feature descriptor, the classifier also has great influence on the performance of human detection. The Support Vector Machine (SVM) and variants of boosted decision tree are two leading classifiers due to their good efficiency [5]. Oren *et al.* [8] firstly introduced machine-learning technology into human detection. They used SVM to train human detector, which was frequently adopted [1, 5, 18]. However, high dimensional features were needed for guaranteeing detection performance, which were time-consuming in sliding-window based detection system.

To improve processing speed, many approaches have been proposed [2, 4, 12, 19-22]. One of the most important methods was proposed by Viola *et al.* [4]. They used Haar-like features and Adaboost to train cascaded classifier for face detection. With the help of simple features, the integral image technology and the cascaded structure of classifier, this method achieved real-time speed with good detection performance. Inspired by [1] and [4], Zhu *et al.* [2] trained a cascaded classifier by AdaBoost based on variable-sized and block-based HOGs. Their cascaded classifier was proved to be dozens of times faster than the SVM classifier in [1]. In recent years, the idea of cascading different kinds of classifiers was also proposed to improve human detection performance [3, 14]. Zeng *et al.* [3] used mi-SVM (Support Vector Machine for multiple instance learning) to train the HOG and LBP feature respectively, and then cascaded the two mi-SVM classifiers directly. How to select and combine different kinds of classifiers and construct cascaded rejecters are important issues in designing human detection system for real applications.

In this paper, inspired by some present research [1, 2, 3, 5], based on 36 dimensional block-based HOGs [1, 2], we propose a feature selection and combination framework to retain discriminative local information and global information of human object. Besides, for real applications, we propose a method to cascade different kinds of classifiers to gain robust detection performance with fast speed. The main work in this paper is listed as following:

- 1). We present a set of Related HOG (RHOG) features including Elementary HOGs (Ele-HOGs) and Concatenation of Ele-HOGs (CSele-HOG). Ele-HOGs are discriminative block-based HOGs [1, 2] selected by AdaBoost, which describe local

distribution of human object. CSele-HOG is a vector concatenated by Ele-HOGs, which contains global and correlative information.

2). Based on RHOG features, we propose a novel human detection scheme using Cascaded Adaboost and SVM classifiers (CAS). Firstly, Adaboost is used to select Ele-HOGs from a huge number of block-based HOGs. The first several stages of Adaboost are used as the first part of our CAS scheme. This part can reject most non-human candidates quickly. Secondly, Ele-HOGs are concatenated to be CSele-HOG descriptor and a SVM classifier is trained. This SVM classifier is used as the last part of our CAS scheme, which can guarantee a high detection performance.

3). When RHOG features are used in our CAS scheme, Ele-HOGs discriminatively describe the local information of human object while CSele-HOG describes the global information. So we simultaneously make use of the local and global information of each Ele-HOG. Moreover, it does not need extra time to calculate CSele-HOG because it is the by-product of computing Ele-HOGs. Experimental results on INRIA dataset show that using RHOG features in CAS framework achieves better detection performance compared with the state-of-the-art human detectors [1, 2].

2 Related Work

Dalal *et al.* [1] proposed HOG descriptor which is considered as one of the most successful features for human detection. In this approach, the 36 dimensional block-based HOG can effectively describe local information of human object. However, the extraction of HOGs is restricted to a single square scale (block with size of 16×16 pixels), so some distinctive HOG information within other variable scales or sizes may be omitted. Meanwhile, the HOG descriptor of each scanning window, which is constituted by 105 gradient histograms extracted from $7 \times 15 = 105$ blocks [1], may contain some redundancy information. Finally, the 3780 dimensional HOG descriptor is trained by SVM classifier, which is time-consuming in sliding-window based human detection system.

Zhu *et al.* [2] adopted Adaboost to select distinctive HOGs from a feature pool which is constructed by 5031 variable-sized and block-based HOGs. In this approach, many selected HOGs are big blocks which are not contained in the 105 fixed blocks in [1]. With the help of Adaboost and integral image techniques, this approach obtains faster speed with similar detection performance compared with [1]. However, this approach only uses block-based HOGs, which could well describe local distribution of human object but lost global information. Moreover, each weak classifier is based on 36 dimensional HOGs, which don't use the correlation information between different distinctive HOGs.

Ye *et al.* [3] proposed a two-stage classifiers scheme to combine coarse features and fine features. Adaboost is used to select coarse features in the first stage to guarantee high speed, and SVM is used to train fine features to gain high detection accuracy. In this method, the coarse features are the unit orientations while the fine features are the pixel orientation histograms of the unit. The fine feature had no relationship with the coarse feature.

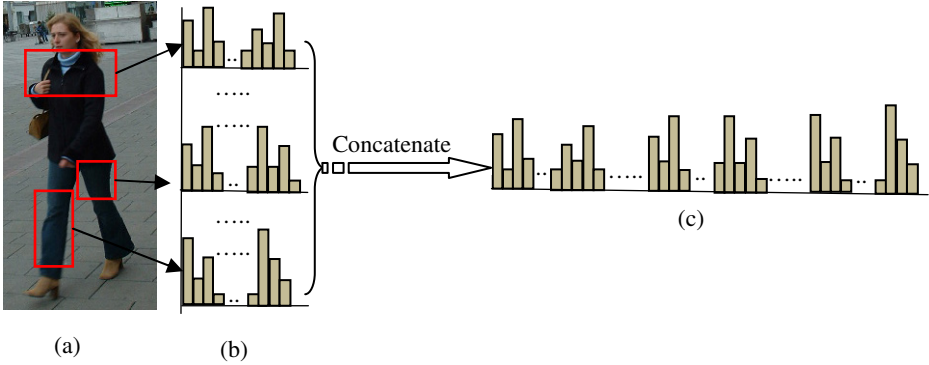


Fig. 1. Description of RHOG features. (a). some selected variable-sized blocks (b). some Ele-HOGs (c). construction of CSele-HOG descriptor.

3 Our Proposed Method

3.1 Related HOG Features (RHOG)

We propose a set of Related Histogram of Oriented Gradients (RHOG) features, which contains Elementary HOGs (Ele-HOGs) and the Concatenation of Ele-HOGs (CSele-HOG). Ele-HOG is a vector of block-based HOG with 36 dimensions [1,2] which are selected by Adaboost from a huge number of candidate HOGs, and CSele-HOG is a vector concatenated by Ele-HOGs.

Ele-HOG Feature. The construction method of Ele-HOG feature is similar to [1, 2]. Firstly, each detection window is divided into variable-sized blocks [2] and each block is divided into 4 cells; then the orientation over $0^{\circ}\sim 180^{\circ}$ is divided into 9 bins. So each cell consists of a 9-bin Histogram of Oriented Gradients (HOG) and each block contains a concatenated 36 dimensional HOG feature; finally, all the block-based HOGs in the sliding window are used to constitute a feature pool, from which Adaboost is use to select Ele-HOGs, as shown in Figure 1 (b). In order to obtain Ele-HOGs quickly, we use integral images technology [4] and Convolved Trilinear Interpolation (CTI) [5] to replace the trilinear interpolation in approach [1]. The improved Ele-HOG can train a fast cascaded rejecter by Adaboost.

CSele-HOG Descriptor. The CSele-HOG descriptor is concatenated by all Ele-HOGs selected by Adaboost as shown in Figure 1 (c). The dimension of CSele-HOG is decided by the number of Ele-HOGs, which can be calculated by the following equation:

$$D_{concatenation} = N * D_{ele} \tag{1}$$

Where D_{ele} means the dimension of Ele-HOG, which is 36 in this paper. N denotes the number of Ele-HOGs selected by Adaboost. The value of N is decided by the number

of stages of the cascaded rejecter. So we can select a proper number of cascaded stages to control the dimension of CSele-HOG descriptor.

Here, from a huge number of variable-sized and block-based HOGs, some discriminative ones are selected by Adaboost, which are called Ele-HOGs in this paper. These Ele-HOGs describe local distribution of human object discriminatively. In order to use the correlation of these distinctive Ele-HOGs, we concatenate them into a vector named CSele-HOG descriptor. Compared with Ele-HOGs, CSele-HOG descriptor contains discriminative global information, which can be used to train a SVM classifier with high detection performance.

CSele-HOG descriptor is concatenated by Ele-HOGs directly, and they have close relationship with each other. Therefore, we call them RHOG features. In the next section, RHOG features are used in Adaboost cascaded rejecter and SVM classifier of the CAS framework respectively, which can improve detection performance and processing speed at the same time.

3.2 Human Detection Scheme of Cascaded Adaboost and SVM Classifiers (CAS)

Based on RHOG features, we propose a novel human detection scheme of Cascaded Adaboost and SVM classifiers (CAS), aiming at achieving high detection performance and fast speed.

Training. We train the cascaded rejecter and SVM classifier on public data set INRIA [1]. The training data set of INRIA includes 2476 positive human patches (including left-right reflections) and 1218 negative non-human images.

Firstly, we use linear SVM to train weak classifiers of the cascaded rejecter and use Adaboost to select the discriminative weak classifiers to construct strong classifiers. The feature pool contains 2346 variable-sized HOGs, from which 5% HOGs are randomly sampled to train weak classifiers just as [2]. Here, we choose 1238 positive samples (the non-left-right ones) and 4000 negative samples randomly selected from the 1218 non-human images to construct training data set. The negative samples are updated in each round of strong classifier training process.

Secondly, we combine distinctive Ele-HOGs, selected by Adaboost, to construct CSele-HOG descriptor and use it to train SVM classifier with more discriminative global and correlation information. Inspired by [1], we use the above Adaboost cascaded rejecter to choose hard samples for CSele-HOG SVM classifier training. Figure 2 shows the details of the selection process of hard samples and the training process of CSele-HOG SVM classifier. First, as the red trace in Figure 2 shows, we use cascaded rejecter to detect the 2476 positive patches and choose the right detections as initial positive samples, while randomly detect the 1218 negative images and choose the wrong detections as initial negative samples. We use these initial hard samples to train the initial CSele-HOG SVM classifier. Then, as the blue trace in Figure 2 shows, we use the concatenated scheme of Adaboost cascaded rejecter and initial CSele-HOG SVM classifier to select more negative hard samples by exhaustively search false positives in the 1218 negative images. We combine the initial hard samples with the exhaustively searched hard samples to retrain the CSele-HOG SVM classifier to produce the final SVM classifier. The number of positive and negative samples used for the final SVM classifier training is decided by the cascaded

rejecter and the scanning parameters of exhaustively searching which will be discussed in experimental section.

The size of all classifiers used in our experiment is 64×128 pixels. On the PC with 2.93 GHz CPU and 2GB memory, it takes several hours to train CSele-HOG SVM classifier, and several days to train cascaded rejecter by Adaboost. Each stage satisfies minimum detection rate of 0.999 and maximum false positive of 0.5.

Detecting. We use sliding-window technology to detect human object. For real-time applications, the ideal detection system should ensure the detection accuracy and at the same time have high detection speed. In this paper, we propose a novel human detection scheme of Cascaded Adaboost and SVM classifiers (CAS). Firstly we use the cascaded rejecter, which is trained by Adaboost based on Ele-HOG feature, to reject most non-human candidates quickly; then we use SVM classifier, which is based on CSele-HOG descriptor, to guarantee detection accuracy. The details are shown in Figure 3.

In the CAS scheme, we use Ele-HOGs and CSele-HOG descriptor in Adaboost cascaded rejecter and SVM classifier respectively. Ele-HOGs can be extracted very quickly and the cascaded rejecter can guarantee fast speed. CSele-HOG descriptor is concatenated by distinctive Ele-HOGs, so CSele-HOG SVM classifier can make sure high detection accuracy. Furthermore, we do not need to spend extra time to compute CSele-HOG descriptor because it is the by-product of selecting Ele-HOGs.

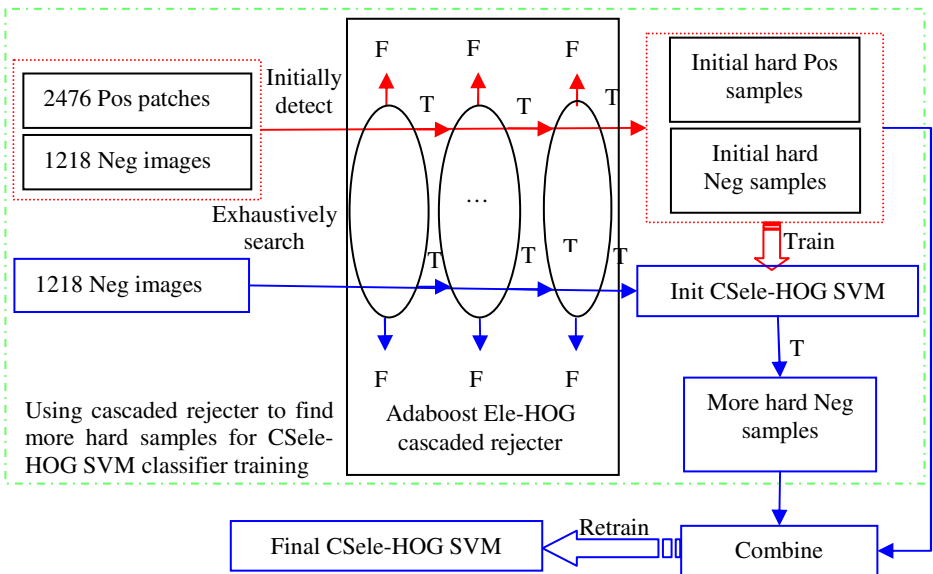


Fig. 2. The training process of the final SVM classifier based on CSele-HOG. (the red trace is the process of selecting initial hard positive and negative samples by using Adaboost, and then using them to train the initial CSele-HOG SVM classifier; the blue trace is the process of selecting more negative hard samples by using cascaded Adaboost and the initial CSele-HOG SVM classifier to exhaustively search the 1218 person-free images, and then combing them with the initial positive and negative hard samples to train the final CSele-HOG SVM classifier).

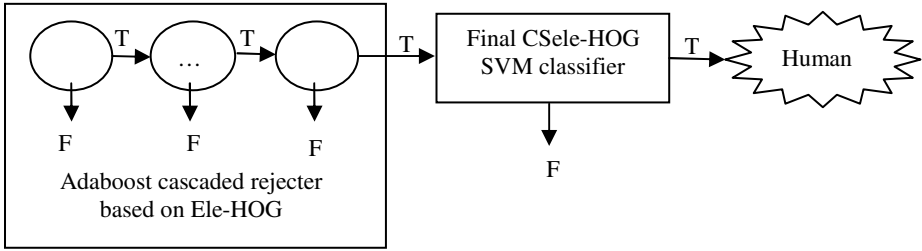


Fig. 3. Proposed human detection scheme based on Cascaded Adaboost and SVM classifier (CAS)

4 Experimental Results

4.1 Introduction of Experimental Condition

To quantitatively analyze classifier performance, we plot Detection Error Tradeoff (DET) curves on a log-log scale, i.e. miss rate (calculated in equation [2]) versus FPPW (false positives per window). Lower values are better. In the experiments, we use miss rate at 10^{-4} FPPW as a reference point for results analyzing.

$$miss\ rate = \frac{False\ Negatives}{True\ Positives + False\ Negatives} \quad (2)$$

The details of training process are described in section 3.2. In our experiments, the test data set comes from INRIA dataset [1], including 1106 human patches with 64×128 pixels and 453 non-human images with variable sizes from 320×240 to 648×748 pixels. We obtain miss rate by using classifiers to detect the 1106 human patches, and get FPPW by using classifiers to scan the 453 non-human images with the scanning parameters as following: scale = 1.12 and stride = (8, 8). The total number of non-human patches is 3,150,775.

4.2 Experimental Results and Analysis

For fair comparison, we train an HOG SVM classifier using approach [1]. The only difference is that we use integral images [4] and the Convolved Trilinear Interpolation (CTI) [5] to replace the trilinear interpolation in [1]. Results show our HOG SVM classifier is 5 times faster than the one provided in OpenCV [1], with similar detection accuracy. In the following experiment, we will compare our method with this improved HOG SVM classifier.

We design three sets of experiments to evaluate our approach. 1) We validate the effectiveness of CSele-HOG descriptor by comparing CSele-HOG SVM classifier with HOG SVM classifier and Zhu’s HOG cascade-of-rejecters [2]. 2) We evaluate the performance of our RHOG-based human detection scheme CAS by comparing CAS scheme with the scheme which only uses CSele-HOG SVM classifier. 3) We evaluate RHOG features by comparing the combination of Ele-HOG cascaded rejecter and CSele-HOG SVM classifier with that of Ele-HOG cascaded rejecter and HOG SVM classifier.

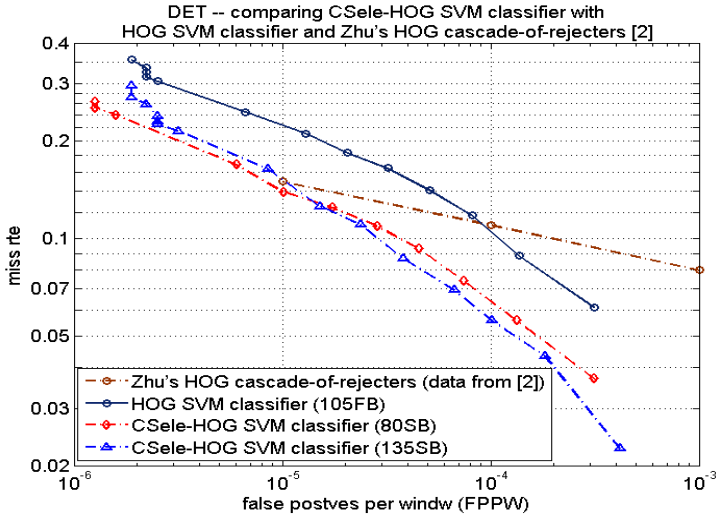


Fig. 4. Comparison of CSele-HOG SVM with HOG SVM [1] and Zhu's [2]

CSele-HOG Feature. To compare with HOG descriptor [1], which is extracted from 105 fixed blocks (105FB), we use Adaboost to select similar number of Ele-HOGs to construct CSele-HOG descriptor. As section 3.2 describes, in our experiment, the number of Ele-HOGs selected by the first 4 and 5 stages of the cascaded rejecter is 80 and 135 respectively. We construct two CSele-HOG descriptors by 80 selected Ele-HOGs and 135 selected Ele-HOGs, and train two SVM classifiers – CSele-HOG SVM (80SB) and CSele-HOG SVM (135SB).

We compare CSele-HOG SVM with HOG SVM and Zhu's HOG cascade-of-rejecter to validating the effectiveness of the proposed CSele-HOG descriptor. The result is shown in Figure 4, among which the data of Zhu's method comes from [2].

Compared with "HOG SVM (105FB)" at 10^{-4} FPPW, our "CSele-HOG SVM (80SB)" and "CSele-HOG SVM (135SB)" improve performance by 4.5% and 5% respectively. Even using fewer blocks, our "CSele-HOG SVM (80SB)" is more discriminative. The reason is that CSele-HOG descriptor is concatenated by distinctive Ele-HOGs which are selected by Adaboost from a huge number of variable-sized and block-based HOGs. However, HOG descriptor is combined by unselected fixed block-based HOGs which may contain several indistinctive ones while miss some distinctive ones of other sized blocks.

Compared with Zhu's method [2], our CSele-HOG SVM also gain better performance at 10^{-4} FPPW. Though [2] use Adaboost to select distinctive HOGs (Ele-HOGs), these features only contain local information and ignore the correlation of these features.

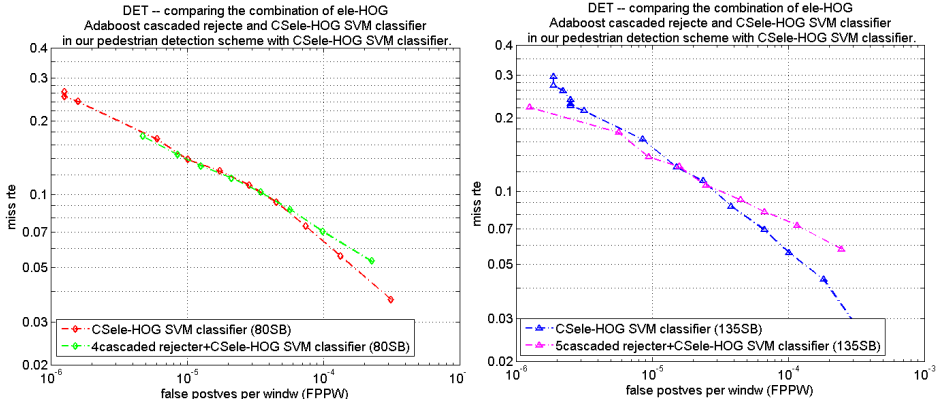


Fig. 5. Results of CSele-HOG SVM and CAS with “cascaded rejecter+CSele-HOG SVM”

CAS Scheme. We test our human detection scheme by comparing CAS scheme which consists of the first 4 stages of Ele-HOG cascaded rejecter and CSele-HOG SVM (80SB) with a scheme which only uses CSele-HOG SVM (80SB), and comparing the CAS scheme that consists of the first 5 stages of Ele-HOG cascaded rejecter and CSele-HOG SVM (135SB) with a scheme that only uses CSele-HOG SVM (135SB).

As Figure 5 illustrates, the detection performance at 10⁻⁴ FPPW decreases a little by adding 4 or 5 stages of cascaded rejecter before SVM classifier both in the two sets of comparisons,. The reason is probably that cascaded rejecter rejects some true positives which will not be rejected by the SVM. Meanwhile, at lower FPPW, the detection performance declines fewer and even increases, because the cascaded rejecter and the SVM classifier may be complementary, resulting in less false positives. Moreover, the detection speed gets faster because cascaded rejecter can reject most non-human candidates at the first several stages quickly, e. g. the first 4

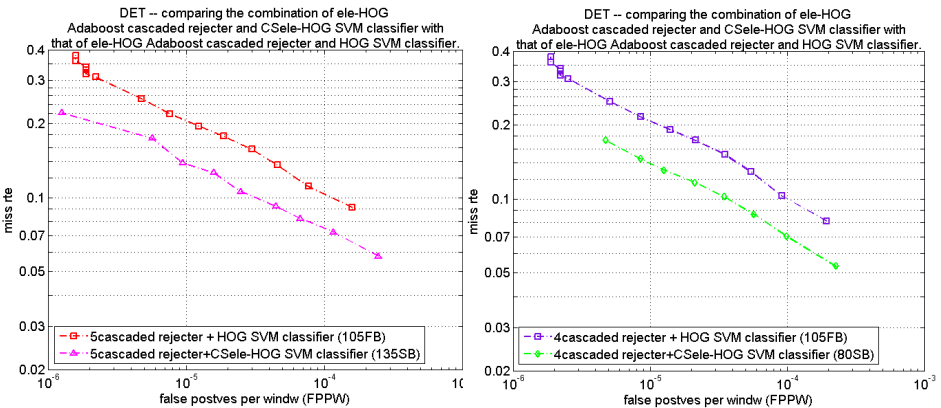


Fig. 6. Comparisons of “cascaded rejecter + CSele-HOG SVM” with “cascaded rejecter + HOG SVM”

stages can reject more than 90 percent of detection windows. Consequently, by adding several stages of cascaded rejecter before SVM classifier, we combine the advantages of Adaboost and SVM methods, achieving faster speed without sacrificing detection performance.

RHOG Features in Our CAS Scheme. We test the performance of RHOG features, by comparing the combination of Ele-HOG cascaded rejecter and CSele-HOG SVM (“cascaded rejecter + CSele-HOG SVM”) with that of Ele-HOG cascaded rejecter and HOG SVM (“cascaded rejecter + HOG SVM”) in CAS scheme.

In the two sets of experiments shown in Figure 6, compared with “cascaded rejecter + HOG SVM”, “cascaded rejecter + CSele-HOG SVM” method improves the detection performance by about 3% at 10^{-4} FPPW while achieves a faster speed. The main reason for the higher performance is that CSele-HOG descriptor is more discriminative than HOG descriptor. The main reason for the faster speed is that as the by-product of selecting Ele-HOGs, CSele-HOG descriptor does not cost extra computing time. Some results of “4cascaded rejecter + CSele-HOG SVM(80SB)” are shown in Figure 7.

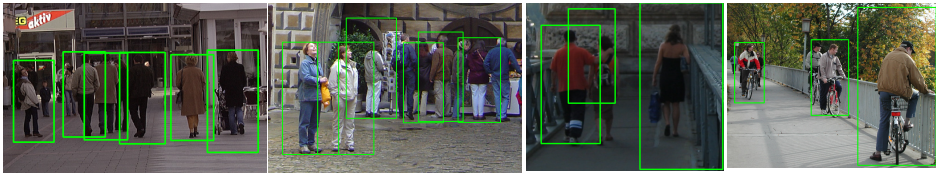


Fig. 7. Some results of our “4 cascaded rejecter + CSele-HOG SVM (80SB)” on INRIA

5 Conclusion

In this paper, we present a set of Related HOG features to describe distinctive local and global information of human object, including Ele-HOGs and a CSele-HOG descriptor. Meanwhile, we propose a novel human detection scheme of Cascaded Adaboost and SVM classifiers (CAS) to combine advantages of Adaboost and SVM. The experimental results show that CSele-HOG descriptor is more discriminative than original HOG descriptor [1]. Moreover, using RHOG features in our CAS scheme can achieve robust and fast detection performance. The Ele-HOG based cascaded rejecter in the proposed CAS scheme can reject most non-human candidates very quickly while the CSele-HOG based SVM classifier can obtain high detection performance.

In the future, we want to combine other features, such as LBP, in our scheme and try to extend the proposed method to handle variations in views. In addition, we will evaluate the proposed method on more public data sets.

Acknowledgments. This work is supported in part by National Nature Science Foundation of China: 60802067, in part by Ningbo Nature Science Foundation: 2012A610046, and in part by Beijing Natural Science Foundation: 4122079.

References

1. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: IEEE Int. Conf. CVPR, pp. 886–893 (2005)
2. Zhu, Q., Yeh, M.-C., Cheng, K.-T., Avidan, S.: Fast human detection using a cascade of histograms of oriented gradients. In: IEEE Int. Conf. CVPR, pp. 1491–1498 (2006)
3. Ye, Q.X., Jiao, J.B., Zhang, B.C.: Fast Human object Detection with Multi-scale Orientation Features and Two-Stage Classifiers. In: IEEE Int. Conf. ICIP, pp. 881–884 (2010)
4. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Int. Conf. CVPR (2001)
5. Wang, X., Han, T.X., Yan, S.: An HOG-LBP Human Detector with Partial Occlusion Handling. In: IEEE Int. Conf. ICCV, pp. 808–820 (2009)
6. Sabzmejdani, P., Mori, G.: Detecting Pedestrians by Learning Shapelet Features. In: IEEE Int. Conf. CVPR, pp. 1–8 (2007)
7. Enzweiler, M., Gavrilu, D.M.: Monocular pedestrian detection: Survey and experiments. IEEE Trans. on PAMI 31(12), 2179–2195 (2009)
8. Oren, M., Papageoriou, C., Sinha, P., Osuna, E., Poggio, T.: Pedestrian Detection Using Wavelet Templates. In: IEEE Int. Conf. CVPR, pp. 193–199 (1997)
9. Gavrilu, D.M., Philomin, V.: Real-Time Object Detection for “Smart” Vehicles. In: IEEE Int. Conf. ICCV, pp. 87–93 (1999)
10. Duan, G., Huang, C., Ai, H., Lao, S.: Boosting associated pairing comparison features for pedestrian detection. In: IEEE Visual Surveillance Workshop (2009)
11. Papageorgiou, C., Poggio, T.: A trainable system for object detection. International Journal of Computer Vision 38(1), 5–33 (2000)
12. Wu, B., Nevatia, R.: Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors. In: IEEE Int. Conf. ICCV, pp. 90–97 (2005)
13. Viola, P., Jones, M.J., Snow, D.: Detecting Pedestrians Using Patterns of Motion and Appearance. In: IEEE Int. Conf. ICCV, pp. 734–741 (2003)
14. Zeng, C.B., Ma, H.D., Ming, A.L.: Fast Human Detection Using MI_SVM and a Cascade of HOG-LBP Features. In: IEEE Int. Conf. ICIP, pp. 3845–3848 (2010)
15. Zhang, J.G., Huang, K.Q., Yu, Y.N., Tan, T.N.: Boosted Local Structured HOG-LBP for Object Localization. In: IEEE Int. Conf. CVPR (2011)
16. Watanabe, T., Ito, S., Yokoi, K.: Co-occurrence Histograms of Oriented Gradients for Pedestrian Detection. In: Wada, T., Huang, F., Lin, S. (eds.) PSIVT 2009. LNCS, vol. 5414, pp. 37–47. Springer, Heidelberg (2009)
17. Pang, Y.W., Yan, H., Yuan, Y.: Robust CoHOG Feature Extraction in Human Centered Image/Video Management System. IEEE Trans. on Systems, Man, and Cybernetics 42(2), 458–468 (2012)
18. Maji, S., Berg, A.C., Malik, J.: Classification Using Intersection Kernel SVM is Efficient. In: IEEE Int. Conf. CVPR, pp. 1–8 (2008)
19. Hou, C., Ai, H., Lao, S.: Multiview Pedestrian Detection Based on Vector Boosting. In: Asian Conference on Computer Vision, pp. 18–22 (2007)
20. Jia, H.X., Zhang, Y.J.: Fast Human Detection by Boosting Histograms of Oriented Gradients. In: Int. Conf. on Image and Graphics, ICIG, pp. 683–688 (2007)
21. Xu, T., Liu, H., Qian, Y.L., Wang, Z.: A fast and robust pedestrian detection framework based on static and dynamic information. In: IEEE Int. Conf. ICME, pp. 242–247 (2012)
22. Laptev, I.: Improvements of Object Detection Using Boosted Histograms. In: Proc. BMVC, Edinburgh, UK, pp. 949–958 (2006)

Face Recognition Using Multi-scale ICA Texture Pattern and Farthest Prototype Representation Classification

Meng Wu, Jun Zhou, and Jun Sun

Institute of Image Communication and Information Processing,
Shanghai Jiao Tong University, Shanghai, 200240, China
wmeng@sjtu.edu.cn

Abstract. In this paper, we present a novel approach to improve the performance of face recognition. To represent face images, we propose an effective texture descriptor, i.e., multi-scale ICA texture pattern (MITP). MITP generates multiple encoded images according to the order of response images by learned independent component analysis (ICA) filters of various scales, and then concatenates the MITP histograms from non-overlapping subregions of the encoded images into a single histogram. Based on a fundamental concept that a specific class can be modeled by a single query-dependent prototype, we introduce a simple classifier without parameter tuning, in which the decision is made using the farthest prototype rule. Moreover, a simple feature remapping strategy can further boost the performance. Experiments on two widely-used face databases demonstrate the effectiveness of our approach over other methods.

Keywords: face recognition, multi-scale ICA texture pattern (MITP), farthest prototype rule, feature remapping.

1 Introduction

Nowadays, tons of social media data are being conveyed and shared through the web, including social networking websites (e.g., Facebook), photo and video sharing websites (e.g., Flickr, Youtube), etc. Personal photographs captured in digital form are increasingly a large portion of these social media data. Automatic face recognition is vital to the content understanding of these pictures, since it allows photos to be tagged and organized by the identities of the individuals conveniently. Generally speaking, a typical face recognition system consists of the following three stages: face detection, face representation (i.e., representing a face image by an efficient and discriminative feature) and face classification. Face representation usually includes feature design and feature extraction/selection [1]. It's self-evident that the last two stages are crucial for good recognition performance.

For decades, a great many approaches have been proposed from various angles. Some focus on how to extract robust and discriminative features [1–4], while others place emphasis on the design of the classifiers [5, 7–9]. Recently,

local binary pattern (LBP) have been successfully exploited in the facial image analysis tasks including face identification, due to its robustness to monotonic illumination variations and computational simplicity [2]. Jabid et al. [3] proposed a more compact facial descriptor named local directional pattern (LDP), which has gained impressive results in face recognition. Gabor-based feature was reported to greatly increase the recognition rates because Gabor features make it feasible to extract local image directional features at multiple scales (levels) [1, 4]. However, one limitation of these features is that their adopted masks are hand-crafted (e.g., LDP uses eight Kirsch masks.). Besides, Gabor feature is computationally cumbersome, since it usually boasts of a high dimension even after a down-sampling procedure.

In terms of classifier design, Wright et al. [5] proposed a sparse representation based classifier (SRC) stemming from sparse coding mechanism, and the state-of-the-art performance has been obtained in face recognition. But in the practical implementation, an exhaustive search of the optimum sparsity controlling parameter of SRC is not feasible. To avoid the difficulties of parameter selection, many classifiers without tuning any parameter have emerged henceforth from the viewpoint of prototype reduction [6]. Linear regression-based classification (LRC) algorithm adopts least squares to represent a query sample as a linear combination of class-specific training samples and the decision is ruled in favor of the class with the minimum reconstruction error [7]. Utilizing the class mean as the prototype, a mean representation based classifier (MRC) was proposed, in which the class with the largest coefficient is favored [8]. Xu et al. [9] proposed a classifier (denoted as NTSRC) by selecting the nearest training samples (NTS) of the test (query) sample from each class as the class prototypes and the minimal construction residuals using least squares can be used to review the identity. Obviously, NTS and LRC are query dependent but class mean is otherwise. Decisions of all the three classifiers are made by the nearest prototype (subspace) rule. Bearing the aforementioned pros and cons in mind, we propose both a novel texture descriptor and a parameter-free classifier to improve the recognition rates while facilitating the process.

The remainder of this paper is organized as follows. Section 2 introduces the proposed MITP descriptor in detail and the feature remapping strategy. Section 3 presents the farthest prototype representation classification algorithm. Experimental results and conclusion are presented in Section 4 and 5, respectively.

2 Multi-scale ICA Texture Pattern (MITP)

In this section, we expect to create a compact and discriminative feature inheriting the merits of previous features. In terms of the definition of the convolution masks, we attempt to vary both the sizes and the coefficients of the masks, thereby obtaining a group of learned filters (basis images) for face representation. To adaptively capture such intrinsic variations, we learn the filters from a bunch of textured patches which are randomly sampled from the training images. Assuming these patches follow a certain distribution, we anticipate

an appropriate model to describe this distribution. Principal component analysis (PCA) regards patches as random variables with the Gaussian distribution and only minimizes the second-order statistics. However, for any non-Gaussian distribution especially in the application of face image analysis, the largest variance would not correspond to PCA basis vector. To counteract this, we resort to independent component analysis (ICA) [10] for learning data dependent filters, which takes both second-order and higher-order statistics into account and is more capable of capturing detailed information. The model of ICA is given by

$$\mathbf{x}_p = \sum_{i=1}^N s_i \mathbf{a}_i = \mathbf{A} \mathbf{s} \quad (1)$$

where \mathbf{x}_p is a vectorized patch, $\mathbf{s} = [s_1, s_2, \dots, s_N]^T$ is the component vector, and the basis images \mathbf{a}_i are the columns of matrix \mathbf{A} . ICA tries to estimate the statistically independent components s_i by computing the (pseudo) inverse of \mathbf{A} , say \mathbf{W} . The row vectors in \mathbf{W} are the learned basis images, i.e., the ICA filters we anticipate. Obviously, the number of the basis images is controlled by the size of the sampled patches. PCA is often used to reduce the dimension before performing ICA, which will accordingly determine the eventual number of basis images. Some examples of ICA basis images learned from two face databases are shown in Figure 1. It can be observed that the basis images are frequency and orientation selective. In this paper, we only retrieve the first eight basis images, i.e., $M = 8$.

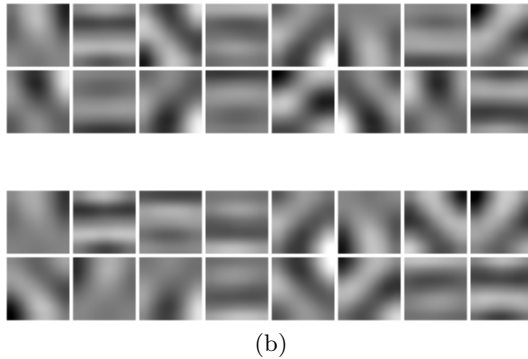


Fig. 1. First eight multi-scale ICA basis images learned from: (a) AR database; (b) Extended Yale B database (Top row and bottom row correspond to sizes 5×5 and 7×7 , respectively.)

Given M learned filters \mathbf{R}_l^m of size $l \times l$, M response images are obtained by convolving a raw image with these filters. To generate a compact feature while remaining the discriminative information, we adopt a similar way as LDP [4]. We sort all the corresponding response images \mathbf{O}_l^m at the same pixel location and

select K ($K = 3$ in this paper) most prominent responses to form the encoded image \mathbf{E}_l , which is expressed in decimal form as

$$\mathbf{E}_l(i, j) = \sum_{m=0}^{M-1} \delta(\mathbf{O}_l^m(i, j))2^m, \tag{2}$$

$$\delta(x) = \begin{cases} 1 & x \in \mathbf{T}_K(i, j) \\ 0 & x \notin \mathbf{T}_K(i, j) \end{cases}$$

where $\mathbf{T}_K(i, j)$ is the set of top K responses at location (i, j) , $\mathbf{O}_l^m = \mathbf{I} \otimes \mathbf{R}_l^m$, and \mathbf{I} is the raw image. It is evident that MITP is compact since it only produces $\frac{M!}{K!(M-K)!}$ different patterns.

For a group of filters of certain size (scale), we can get an encoded image. To capture richer information, ICA filters of P different sizes are used to generate P encoded images. In our work, we merely choose the filters of two scales as shown in Figure 1. Each encoded image is then divided into non-overlapping subregions for economic use and the subregion histograms are concatenated into a compact histogram $\mathbf{H}_l \in \mathbb{R}^q$. Finally, individual histograms at various scales are further concatenated to form the proposed feature $\mathbf{F} \in \mathbb{R}^d$, where $d = Pq$. The whole process of extracting the proposed MITP feature is illustrated in Figure 2.

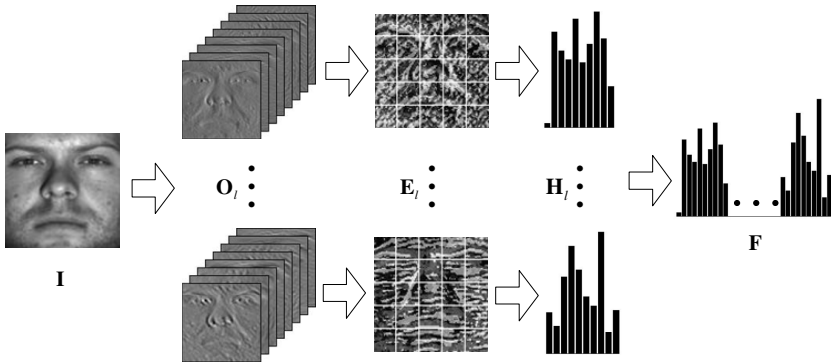


Fig. 2. The whole process of our proposed multi-scale ICA texture pattern (MITP) feature extraction

Since MITP belongs to histogram-based features, a simple α -exponentiation feature remapping $\mathbf{F} \rightarrow \mathbf{F}^\alpha$ with $\alpha < 1$ may improve the performance in view of its success in image classification [11]. We simply adopt the square-rooting feature, i.e., $\alpha = 0.5$, considering that similar results by feature remapping are achieved when α varies in $(0, 1)$. Although square-rooting feature corresponds to an exact mapping of the Bhattacharyya kernel, we regard it as one possible nonlinear remapping of the feature at zero cost.

3 Farthest Prototype Representation Classification

As we have mentioned before, these parameter-free classifiers do not need delicate tuning and have attained satisfying results. The advantages of these classifiers motivate us to develop a simple but effective classifier, called the farthest prototype representation based classifier (FPRC).

Constructing an appropriate prototype model can gain significant speed-up without compromising much accuracy. In our opinion, the query-dependent prototype is more robust than query-independent one since the former is closely related to the query image, thereby capturing more intra-class variances. Therefore, we prefer NTS to class mean for constructing the prototype. Let \mathbf{F}_i^c denote the i -th training sample from the c -th class on the feature space, $i = 1, 2, \dots, n_c$, and $c = 1, 2, \dots, s$. Given a test sample \mathbf{y} , the prototype of NTS is computed by

$$NTS_c = \arg \min_i \|\mathbf{F}_i^c - \mathbf{F}_y\|_2^2 \quad (3)$$

where $n = \sum_{c=1}^s n_c$ is the number of total training samples, and \mathbf{F}_y denotes the MITP feature of \mathbf{y} .

It is assumed that a test sample on the feature space can be approximately represented by a linear combination of all the NTSs, which is formulated as

$$\mathbf{F}_y = \sum_{c=1}^s \beta_c NTS_c = \mathbf{S}\boldsymbol{\beta} \quad (4)$$

where $\mathbf{S} = [NTS_1, \dots, NTS_s]$, and the coefficient vector $\boldsymbol{\beta} = [\beta_1, \dots, \beta_s]^T$. Given that $d \geq s$ holds true in our case (feature dimension is usually larger than the number of classes), Eq.(4) is well conditioned, thus its least squares solution is given by

$$\boldsymbol{\beta} = (\mathbf{S}^T \mathbf{S})^{-1} \mathbf{S}^T \mathbf{F}_y \quad (5)$$

On the testing phase, the nearest prototype rule identifies the face image based on a single class prototype, which is unreliable in our case. To tackle this issue, we resort to other prototypes except a specific class prototype to incorporate more available information. The identity of a test sample is disclosed by the farthest prototype rule considering the fact that there is only one sample for each class, which is expressed as

$$\text{ID}(\mathbf{y}) = \arg \max_c \left\| \sum_{i \neq c} \beta_i NTS_i - \mathbf{F}_y \right\|_2^2 \quad (6)$$

4 Experimental Results

4.1 Databases and Parameter Setting

To fully compare the algorithms, we have conducted the experiments on two benchmark face databases: AR and Extended Yale B (EYaleB) databases [12,13].

For a fair comparison, we duplicate the experimental set-up in [4, 5] for all the experiments and the best recognition rates with the corresponding feature dimension are reported. The AR database consists of over 4,000 images of 126 individuals. For each individual, 26 images were taken in two separate sessions. We choose a subset of the database consisting of 50 male individuals and 50 female individuals. For each subject, the seven images from Session 1 are selected for training, and the other seven images from Session 2 are used for testing. The Extended Yale B database contains 2,414 frontal face images of 38 subjects, captured under various controlled lighting conditions. For each subject, we randomly select half of the images for training (i.e., about 32 images per subject) and the other half for testing.

In the experiments, we will validate each part of our approach by comparing performance of various combinations of different types of features and the classifiers. Specifically, we compare our MITP feature with four related features, i.e., raw pixel, Gabor, LBP and LDP, while we compare the proposed classifier FPRC with three related classifiers: LRC [7], MRC [8] and NTSRC [9]. The subregion division is 5×5 for images on AR and Extended Yale B databases and principal component analysis (PCA) serves as the dimensionality reduction tool. Experiments on the Extended Yale B database are repeated 10 times with random split of training and testing sets.

4.2 Results and Discussion

Tables 1 and 2 tabulate the best recognition accuracy with the corresponding feature dimension for various features, in conjunction with four different classifiers: LRC, MRC, NTSRC, and FPRC on the AR and Extended Yale B databases. It can be clearly observed that our MITP feature performs best among all the features using any classifier (MRC in particular). We attribute this largely to the adaptively learned ICA masks that enables MITP to capture the invariances within a class as much as possible. As to classifier, MRC yields rather unsatisfying results especially in the case of huge intra-class variations. Because the prototype of class mean is query-independent, it fails to reflect large illumination changes on Extended Yale B. Although LRC can do well on Extended Yale B in the case that there exist enough samples per class, it performs much worse than NTSRC and FPRC on AR. NTSRC and FPRC are sparse classifiers in some sense for they discard many irrelevant samples beforehand. FPRC is a bit superior to NTSRC in most cases, which verifies the effectiveness of farthest prototype rule. Note that the proposed MITP feature using FPRC achieves the highest recognition accuracy among all the various combinations on either database.

Table 3 shows the results of our proposed approach by feature remapping on two databases. Overall, our method achieves 99.86% accuracy on AR and 99.20% accuracy on Extended Yale B, which validates that the feature remapping strategy does ameliorate the results.

Table 1. Best recognition accuracy with the corresponding feature dimension for various combinations on the AR database (%)

Feature	Classifier			
	LRC	MRC	NTSRC	FPRC
Raw pixel	76.68 (450)	71.82 (550)	84.55 (500)	83.12 (600)
Gabor	83.55 (500)	87.12 (500)	92.56 (600)	94.85 (600)
LBP	85.84 (200)	82.40 (500)	94.99 (400)	95.57 (600)
LDP	89.56 (150)	54.08 (400)	87.70 (550)	81.55 (550)
MITP	91.85 (250)	95.28 (550)	98.28 (400)	98.71 (300)

Table 2. Best recognition accuracy with the corresponding feature dimension for various combinations on the AR database (%)

Feature	Classifier			
	LRC	MRC	NTSRC	FPRC
Raw pixel	95.14 ± 0.23 (400)	47.64 ± 2.67 (600)	94.00 ± 0.50 (600)	94.29 ± 0.48 (600)
Gabor	90.82 ± 0.38 (600)	57.59 ± 1.97 (600)	90.39 ± 0.44 (600)	91.60 ± 0.59 (600)
LBP	91.25 ± 0.44 (600)	50.12 ± 3.32 (600)	91.63 ± 0.47 (600)	92.17 ± 0.38 (600)
LDP	96.52 ± 0.57 (300)	52.83 ± 2.69 (600)	95.06 ± 0.63 (450)	95.13 ± 0.69 (550)
MITP	97.63 ± 0.62 (500)	80.84 ± 2.98 (600)	98.06 ± 0.26 (600)	98.60 ± 0.21 (600)

Table 3. Best recognition accuracy with the corresponding feature dimension for MITP + FPRC using feature remapping (%)

Dataset	Accuracy	Dimension
AR	99.86	500
EYaleB	99.20 ± 0.21	200

5 Conclusion

We propose a novel approach for face recognition from the perspective of face representation and classifier design. First, we propose a distribution-based feature by applying the learned ICA filters of various scales and encoding the order of the response images. Second, we incorporate the farthest decision rule into the prototype-representation-based classifier to improve the recognition rates while reducing the computational complexity. Third, the feature remapping strategy is exploited to further improve the effectiveness of MITP feature. Extensive experiments conducted on the AR and Extended Yale B face databases indicate the superiority of our proposed MITP feature and FPRC classifier.

Acknowledgements. This work was supported by MOST project under Contact 2011BAH08B01, and Science and Technology Commission of Shanghai Municipality grant under Contact 10511501102.

References

1. Xie, S.F., Shan, S.G., Chen, X.L., Chen, J.: Fusing local patterns of gabor magnitude and phase for face recognition. *IEEE Trans. Image Process.* 19, 1349–1361 (2010)
2. Huang, D., Shan, C.F., Ardabilian, M., Wang, Y.H., Chen, L.M.: Local binary patterns and its application to facial image analysis: A survey. *IEEE Trans. Syst. Man Cybern. Part C Appl. Rev.* 41, 765–781 (2011)
3. Jabid, T., Kabir, M.H., Chae, O.: Local directional pattern (LDP) for face recognition. *Int. J. Innov. Comput.* 8, 2423–2437 (2012)
4. Yang, M., Zhang, L.: Gabor Feature Based Sparse Representation for Face Recognition with Gabor Occlusion Dictionary. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010, Part VI. LNCS*, vol. 6316, pp. 448–461. Springer, Heidelberg (2010)
5. Wright, J., Yang, A.Y., Ganesh, A., Sastry, S.S., Ma, Y.: Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 210–227 (2009)
6. Triguero, I., Derrac, J., Garcia, S., Herrera, F.: A taxonomy and experimental study on prototype generation for nearest neighbor classification. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.* 42, 86–100 (2012)
7. Naseem, I., Togneri, R., Bennamoun, M.: Linear regression for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 32, 2106–2112 (2010)
8. Xu, J., Yang, J.: Mean representation based classifier with its applications. *Electron. Lett.* 47, 1024–1026 (2011)
9. Xu, Y., Zhu, Q.: A simple and fast representation-based face recognition method. *Neural Comput. Appl.* (2012), doi:10.1007/s00521-012-0833-5
10. Jenssen, R., Eltoft, T.: Independent component analysis for texture segmentation. *Pattern Recogn.* 36, 2301–2315 (2003)
11. Chapelle, O., Haffner, P., Vapnik, V.N.: Support vector machines for histogram-based image classification. *IEEE Trans. Neural Networks* 10, 1055–1064 (1999)
12. Martinez, A.M., Benavente, R.: The AR face database. *CVC Technical Report 24* (1998)
13. Georgiades, A.S., Belhumeur, P.N., Kriegman, D.J.: From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.* 23, 643–660 (2004)

Detection of Biased Broadcast Sports Video Highlights by Attribute-Based Tweets Analysis

Takashi Kobayashi^{1,*}, Tomokazu Takahashi², Daisuke Deguchi³, Ichiro Ide¹,
and Hiroshi Murase¹

¹ Graduate School of Information Science, Nagoya University, Japan
{ide,murase}@is.nagoya-u.ac.jp

² Faculty of Economics and Information, Gifu Shotoku Gakuen University, Japan
ttakahashi@gifu.shotoku.ac.jp

³ Information and Communications Headquarters, Nagoya University, Japan
ddeguchi@is.nagoya-u.ac.jp

Abstract. We propose a method for detecting biased-highlights in a broadcast sports video according to viewers' attributes obtained from a large number of tweets. Recently, Twitter is widely used to make real-time play-by-play comments on TV programs, especially on sports games. This trend enables us to effectively acquire the viewers' interests in a large mass. In order to make use of such tweets for highlight detection in broadcast sports video, the proposed method first performs an attribute analysis on the set of tweets issued by each user to classify which team he/she supports. It then detects biased-highlights by referring to the number of tweets made by viewers with a specific attribute.

Keywords: Twitter, broadcast sports video, highlight detection, play-by-play comments.

1 Introduction

Today, due to the enormous amount of programs broadcast on TV, video summarization techniques are needed. Many methods have been proposed on summarizing various types of broadcast videos, such as news [1], sports [2], and cooking [3]. In these works, videos were mostly summarized based only on the audio-visual information that could be extracted from the video contents themselves. This approach is simple, but its output does not always match a viewer's interest, for example, a viewers' interest in a sports game may only be on their favorite team. This drawback was mostly due to the difficulty in establishing a general framework to obtain information on such interests only from the video content itself.

On the other hand, a micro-blogging service "Twitter¹" is rapidly growing the number of its users. A post on Twitter "tweet" consists of a user name, a

* Presently with Canon Inc.

¹ <http://twitter.com/>

comment (maximum 140 characters), and a time stamp. It is said, as of 2012, over 100 million people use this service as a real-time communication tool because of its ease of use. Recently, as one style of Twitter usage, tweeting while watching TV is becoming popular. This enables us to exchange play-by-play comments on contents of a TV program in real-time with many other users sharing the experience while watching the same program. In case of popular programs, tens of thousands of tweets are posted during the broadcast. These tweets reflect the viewers' interests, opinions, and comments to the TV program and its contents.

Our goal is the summarization of TV programs from the viewers' viewpoints. Aiming at this goal, in this paper, we propose a method for biased-highlights detection from a broadcast sports video referring to information on user attributes obtained by analyzing their tweets. We considered that compared to other kinds of video contents, in case of team sports, viewers' interests are relatively simple; which of the two teams they support. The proposed method first performs an attribute analysis on the set of tweets issued by each user to classify which team he/she supports. It then detects biased-highlights by referring to the number of tweets made by viewers with a specific attribute. The benefit of the viewer classification is that the proposed method can provide different sets of highlights biased by supporters of each team, whereas using all tweets without the viewer classification only provides a set of highlights for both teams.

2 Related Work

Miyamori et al. [4] have proposed a broadcast video summarization method based on viewer's perspectives posted on a live chat forum in a Japanese BBS; 2-channel². However, a live chat forum is not sufficient to obtain interests of general viewers in a large community because only limited users in a small community participate.

On the other hand, tweets posted on Twitter are recently focused as a resource to obtain the viewers' interests on a TV program. There is already a commercial service that visualizes tweets posted on current TV programs and ranks them in real-time according to their popularity. Shamma et al. [5] analyzed the viewers' attention of a TV debate from the contents and the number of tweets during the broadcast. In order to summarize broadcast TV videos, only few researches [6, 7] made use of Twitter up to now, but even they do not consider the difference of viewers' interests.

3 Proposed Method

In this paper, we focus on video highlight detection of team-sports where two teams participate in a game, such as baseball and soccer. Therefore, we expect to obtain two different sets of highlight scenes biased according to the viewer's interest; which team he/she supports.

² <http://2ch.net/>

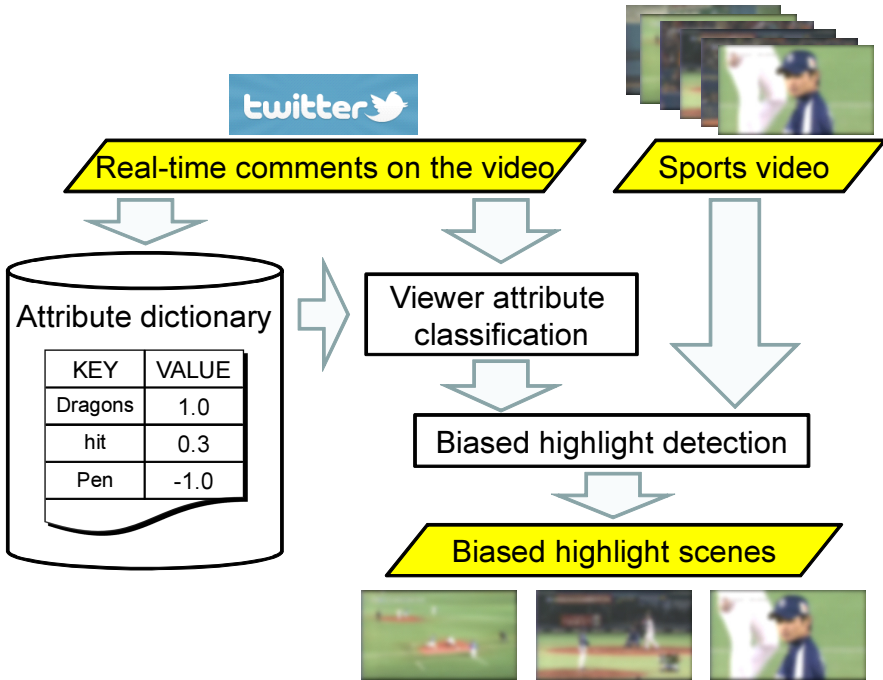


Fig. 1. The flow diagram of the proposed method

Below, we explain the proposed algorithm in the case of a game between teams *A* and *B*. Before the main process, a preprocessing is applied to tweets obtained during the period of the broadcast, in order to filter-out real-time comments by Twitter users who are presumably not watching the program.

Figure 1 shows the flow diagram of the proposed method. The proposed method detects biased-highlights of the video by an attribute-based analysis of the tweets, by mainly the following three steps:

- Creation of an attribute dictionary
- Viewer attribute classification based on their tweets
- Biased highlight detection

Below, we describe the above process in detail.

3.1 Preprocessing

First, tweets related to the actual game are extracted by using their time stamps, keywords including team names, player names, and hash-tags. We consider that the Twitter users whose tweets were included in the extracted tweets are candidates of viewers. Then the bag-of-words obtained from the set of tweets issued by each user is input to a SVM (Support Vector Machine) classifier that is trained to classify viewers and non-viewers. The classifier is trained from tweets on other games.

3.2 Creation of an Attribute Dictionary

This process creates an attribute dictionary that is needed to classify the attribute of each viewer; which team he/she supports. The dictionary consists of pairs of a term and its attribute value. The attribute value is defined in proportion to the frequency of the corresponding term that appeared in tweets by viewers supporting one of the two teams.

Here, the dictionary is renewed per short time period in the video that is being analyzed, because the meanings of a term greatly change according to the context of each game and the plays that occur in it. To create this, we make use of the SO-PMI (Semantic Orientation from Pointwise Mutual Information) method [8], which is a method for unsupervised learning of semantic orientation of a phrase. In this method, the authors classified reviews of products (recommended / not recommended), using semantic orientations of phrases in the reviews that were learned with only two initial given terms; “excellent” for positive orientation and “poor” for negative orientation.

Meanwhile, Twitter users sometimes include hash-tags in the form of “#topic” in their tweets for the purpose of informing other users the topic of their tweets. In case of sports, team names that they support are very often used as hash-tags. Therefore, the hash-tags including the team names were used as the initial terms for the SO-PMI method for the creation of the dictionary in our method. Here, we set the attribute values of hash-tags including team names of A and B to 1 and -1 , respectively. The values of other terms w are calculated by the following equations:

$$V_A(w) = \frac{F_A(w) - F_B(w)}{F_A(w) + F_B(w)} \quad (1)$$

$$F_A(w) = \sum_{T_A \in D_{t,s}} W_{T_A}(w) \quad (2)$$

$$W_{T_A}(w) = \begin{cases} 1 & w \in T_A \\ 0 & \text{otherwise} \end{cases}, \quad (3)$$

where T_A represents a tweet including team A 's hash-tags, and $D_{t,s}$, the set of all the viewers' tweets posted in a short time period (s seconds starting from time t). The value $F_B(w)$ is calculated by replacing A of Eqs. (2) and (3) with B . Thus, a term with positive values ($V_A(w) > 0$) could be used for a term supporting team A , and that with negative values ($V_A(w) < 0$), for team B . Although there are many common terms used to support both teams, the values of these terms become small by Eq. (1).

3.3 Viewer Classification Based on the Attribute Analysis on Their Tweets

Using the attribute dictionary, the attribute of each viewer is classified according to the set of their tweets during a certain period. The attribute of each viewer u is judged by the following equations:

$$L(u) = \begin{cases} A & N_A(u) > 0 \\ B & N_A(u) < 0 \\ Neutral & N_A(u) = 0 \end{cases} \quad (4)$$

$$N_A(u) = \sum_{T_u \in D} \text{sign} \sum_{w \in T_u} V_A(w), \quad (5)$$

where T_u represents a tweet by viewer u , and D , the set of all the viewers' tweets posted in the game. The viewers that post tweets including terms with positive values ($V_A(w) > 0$) are classified as team A 's supporters.

3.4 Biased Highlight Detection

Based on the viewer attribute classification result, this process first divides the tweets into two sets. Each set contains tweets made by viewers labeled as supporting the same team. For each set of tweets, this process then finds the local maxima of the temporal change of the number of the tweets during each short time range as candidates of highlights. The biased highlight scenes are then detected as video segments around the candidates where the number of the tweets is greater than a threshold.

4 Experiment

We applied the proposed method to the fifth game of the annual Japanese baseball championships between “Chunichi Dragons (Hereafter, Dragons)” and “Lotte Marines (Hereafter, Marines)” on November 4, 2011. A total of 20,524 real-time tweets made by 1,424 viewers were obtained after the preprocessing of the proposed method introduced in section 3.1.

4.1 Highlight Detection Accuracy

To evaluate the detection accuracy, we compared the results obtained by the proposed method with the actual highlight scenes that were edited and broadcasted by a local broadcasting station supporting one of the teams; Dragons, their local team.

Figure 2 shows the highlights biased towards the Dragons supporters' viewpoints that were detected by the proposed method. The horizontal axis represents elapsed times from the play-ball, and the vertical axis represents the number of tweets that were posted around the time by users classified as Dragons' supporters. The blue circles and the red triangles represent the biased highlights detected by the proposed method and the broadcasted highlights, respectively. Two of three actually broadcasted highlights were successfully detected by the proposed method when a certain value was used for the detection threshold. Although we determined the threshold manually in this experiment, we will develop a method for automatic determination of the threshold in the future.

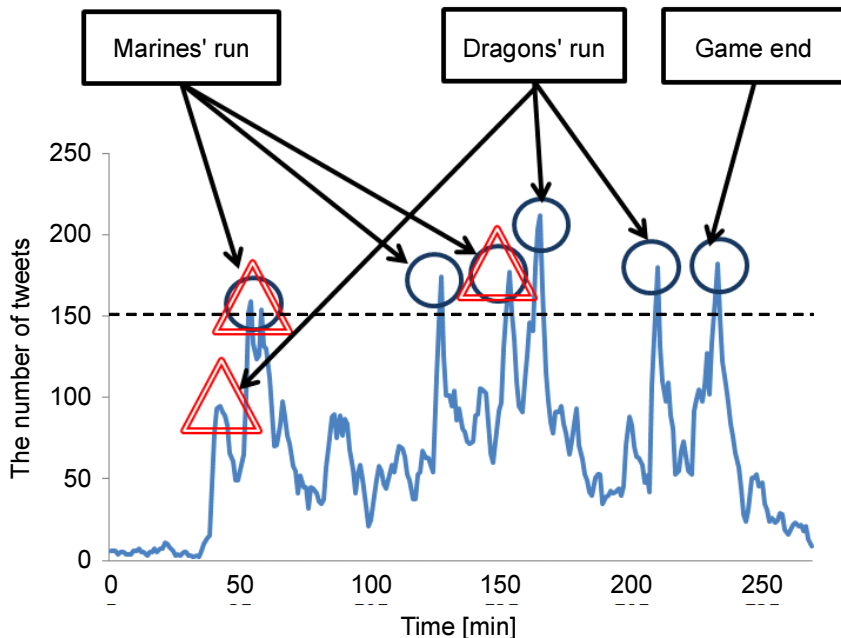


Fig. 2. The highlights biased towards the “Dragons” supporters’ viewpoints obtained by the proposed method

Tables 1 and 2 describe the events that occurred around these moments. Six of the seven detected highlights corresponded to the actual runs in the game. From this, we can confirm that the proposed method successfully detected the highlights. However, the proposed method could not detect the first highlight scene “Dragons score the first run on a sacrifice fly.” As one of the main reasons for this, we consider that the number of tweets around the time was fewer compared to that for other highlights. This is because, for this game, not all TV stations started broadcasting immediately after play-ball. Therefore, only a limited number of viewers that were either on-site, or had access to CATV or satellite TVs could post tweets around the time. To detect highlights accurately including such cases, a method is required that uses not only the number of tweets but also its means and variations.

4.2 Analysis of Viewers’ Interests

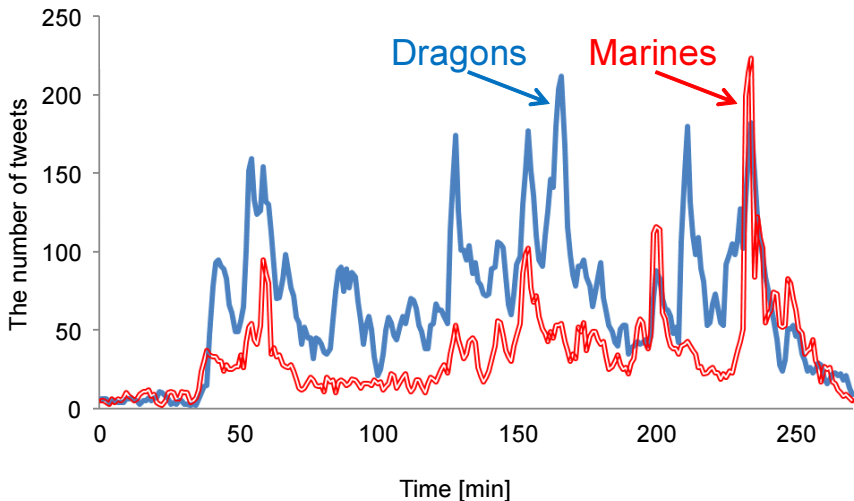
We analyzed the difference of the viewers’ viewpoints depending on the teams they support. As the result of the viewer attribute classification based on the attribute of their tweets, the 1,424 viewers were classified into 684 Dragons’ supporters and 740 Marines’ supporters. Figure 3 compares the transitions of the frequency of tweets issued by the viewers with each attribute. Several peaks where the tweets largely increased was observed in the case of Dragons’ supporters.

Table 1. Biased highlights obtained by the proposed method for Dragons' supporters

Time	Event	Score
18:54	Marines hits with the bases full and turns the game.	4 vs. 1
18:59	Marines adds runs.	6 vs. 1
20:08	Marines hits a two-run home-run.	9 vs. 1
20:34	Marines adds a run.	10 vs. 1
20:45	Dragons adds a run.	10 vs. 2
21:31	Dragons hits a two-run home-run.	10 vs. 4
21:54	The game ends in a win for Marines.	10 vs. 4

Table 2. Highlights broadcasted by a local TV station supporting the Dragons

Time	Event	Score
18:39	Dragons scores the first run on a sacrifice fly.	0 vs. 1
18:54	Marines hits with the bases full and turns the game.	4 vs. 1
20:34	Marines adds a run.	10 vs. 1

**Fig. 3.** The transitions of the frequency of tweets made by the viewers with each attribute

On the other hand, in the case of Marines' supporters, fewer peaks than the Dragons' supporters were observed, with many tweets observed towards the end of the game.

Figures 4 and 5 show the events that occurred around the peaks for each team. We observed that the figures show the difference of interests between the supporters of each team; the Dragons' supporters were interested in scoring scenes by both teams, whereas the Marines' supporters showed interest in scoring scenes of only their team. The reasoning of this difference could be analyzed

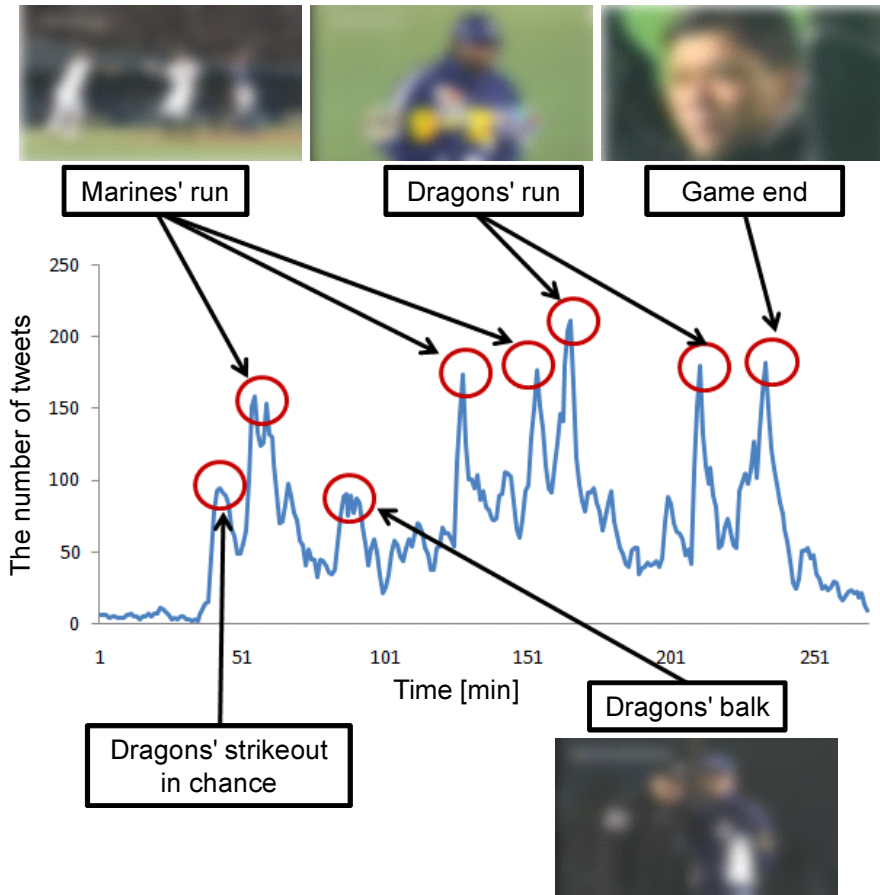


Fig. 4. Interests of the Dragons' supporters

as follows; the game resulted in that Marines won the game by 10 vs. 4 after keeping a large lead for a long time. The Marines' supporters were secure to their victory in the early stage, so they were probably interested only in the Marines' runs that strengthened their feeling of security. On the other hand, the Dragons' supporters kept their hopes on their come-from-behind victory until the last moment, so they were interested in all plays that could affect the game. From this analysis, we confirmed the effectiveness of the attribute-based viewer classification for the acquisition of the viewers' interests.

4.3 Viewer Classification Accuracy

The viewer classification accuracy is one of the most important factors that determines the performance of the proposed method. To evaluate the accuracy, we conducted the following experiment.

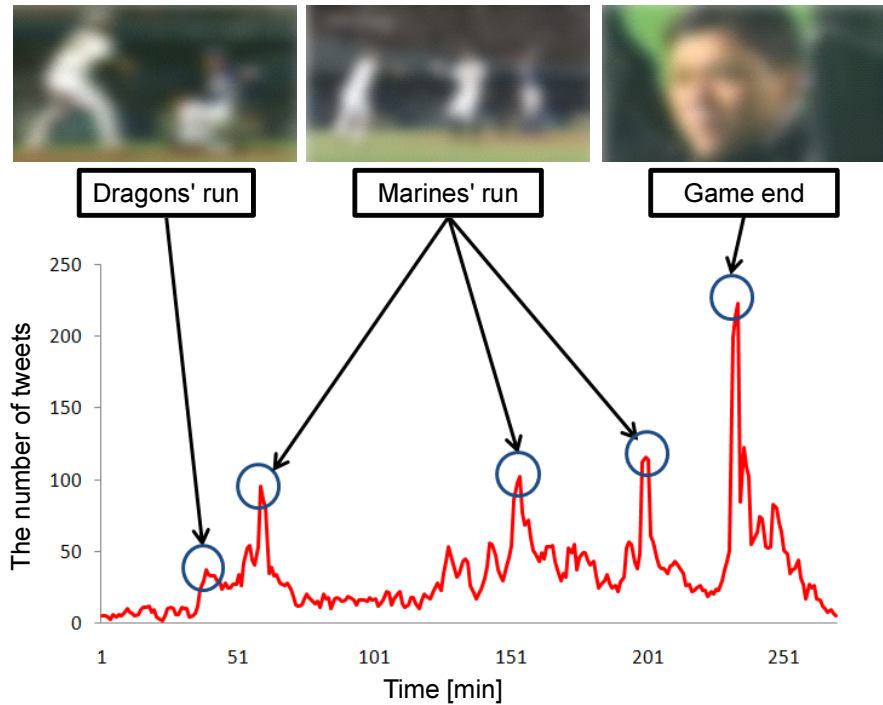


Fig. 5. Interests of the Marines' supporters

First, we chose 200 viewers at random from the 1,424 viewers, and then labeled them manually with their attributes, namely which team he/she supports. The classification accuracy was calculated while changing the unit time length s that was used in the creation of the attribute dictionary. As a result, we confirmed that the proposed method could classify the viewers with the accuracy of more than 80% when s was set to 2 seconds. Shortening s improved the accuracy because the tweets effectively reflected quick responses to each play. On the other hand, too short s sometimes reduced the number of the terms that could be registered to the dictionary, which could cause a degradation in the accuracy.

5 Summary

We proposed a method for biased highlight scene detection from a broadcast sports video based on the interests of viewers obtained through their tweets on Twitter. To acquire each viewer's interest, that is which team the viewer supports, the proposed method performed the viewer classification based on attributes of their tweets. Biased highlights were detected for each team by referring to the transition of the number of tweets by viewers supporting the team. In an experiment, we applied the proposed method to highlight detection

of an actual baseball game broadcasted on TV. From the result, we confirmed that the proposed method could effectively detect highlights biased towards the viewers' interests.

Acknowledgments. Parts of this work were supported by the Grants-in-Aid for Scientific Research.

References

1. Nack, F., Ide, I.: Why did the Prime Minister resign? —generation of event explanation from large news repositories—. In: Proc. Nineteenth ACM Int. Conf. on Multimedia, pp. 313–322 (2011)
2. Chang, Y.L., Zeng, W., Kamel, I., Alonso, R.: Integrated image and speech analysis for content-based video indexing. In: Proc. Int. Conf. on IEEE Multimedia Computing and Systems 1996, pp. 306–313 (1996)
3. Miura, K., Hamada, R., Ide, I., Sakai, S., Tanaka, H.: Motion based automatic abstraction of cooking videos. In: Proc. ACM Multimedia 2002 Workshop on Multimedia Information Retrieval, pp. 21–29 (2002)
4. Miyamori, H., Nakamura, S., Tanaka, K.: Generation of views of TV content using TV viewers' perspectives expressed in live chats on the web. In: Proc. Thirteenth ACM Int. Conf. on Multimedia, pp. 853–861 (2005)
5. Shamma, D.A., Kennedy, L., Churchill, E.F.: Tweetgeist: Can the twitter timeline reveal the structure of broadcast events? In: Proc. 2010 ACM Conf. on Computer Supported Cooperative Work, pp. 589–594 (2010)
6. Hannon, J., McCarthy, K., Lynch, J., Smyth, B.: Personalized and automatic social summarization of events in video. In: Proc. Fifteenth Int. Conf. on Intelligent User Interfaces, pp. 335–338 (2011)
7. Chakrabarti, D., Punera, K.: Event summarization using tweets. In: Proc. Fifth Int. AAAI Conf. on Weblogs and Social Media, p. 154 (2011)
8. Turney, P.D.: Thumbs up? Thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proc. 40th Annual Meeting of the Association for Computational Linguistics, pp. 417–424 (2002)

Temporal Video Segmentation to Scene Based on Conditional Random Fields^{*}

Su Xu, Bailan Feng, and Bo Xu

Institute of Automation
Chinese Academy of Sciences
Beijing 100190, China
{su.xu,bailan.feng,xubo}@ia.ac.cn

Abstract. In this paper, we propose a novel approach of video segmentation into scenes based on the technique of conditional random fields (CRFs). This approach is built upon the design in which scene segmentation is transformed into a label identification problem by defining three types of shots. To implement our algorithm, three middle-level features including shot difference signal, scene transition graph and audio type are extracted to depict the label properties of each shot, and then CRFs model is employed to identify the labels sequence. The advantage of CRFs model lies in its facility in integrating context information of neighboring shots, which produces accurate results in scene segmentation. The proposed approach is verified by seven types of data covering the most major genres of TV program. Experiments on testing data set yield average 0.88 F-measure, which illustrates that the proposed method can accurately detect most scenes in different genres of programs.

Keywords: conditional random field, Scene Segmentation.

1 Introduction

Automatic scene segmentation of video is an essential prerequisite for a wide range of video manipulation applications, such as video indexing, non-linear browsing, classification etc [1]. Video scene segmentation is a bottom-up process. The smallest physical unit of a video is the shot that is defined as an unbroken sequence of frames recorded from the same camera [1]. A scene can be regarded as a series of shots for which the three properties, event or dramatic incident, setting, and time, are consistent [1]. This term is most of the time used with fictional narrative-driven video content, such as movie, TV-series, cartoon and sitcom.

Several approaches are proposed for the scene segmentation problem. In an early stage, graph-based approaches received significant attention for recognizing scene pattern. In [2], Yeung et al. use pair-wise color histogram similarities between key-frames and time-constrained clustering for building scene transition

^{*} This work was supported by the National Natural Science Foundation of China (Grant No.61202326).

graphs (STG) to represent the scenes. A similar approach is presented in [3], where Ngo et al. improve the process of shots clustering by normalized cut algorithm [4] to construct the STG. In [5], Rasheed et al. construct a weighted undirected graph in which the weights are expressed by color histogram similarities and motion information. This graph is iteratively segmented into scene sub-graphs using normalized cuts. The graph-based methods are very likely to break a scene into a few segments leading to poor precision in scene segmentation. To address this problem, some remarkable approaches are presented. In [6], Yun et al. present a statistical approach, based on selecting an initial set of arbitrary scene boundaries and updating them by a Markov chain Monte Carlo (MCMC) technique. In [7], Chasanis et al. conduct shot grouping by spectral clustering, and then a sequence alignment algorithm is applied on the clustering outcome sequences instead of graph model for identifying the scenes. In work [8], Sakarya et al. use graph partition model to construct a one-dimensional signal that is obtained from the similarity matrix in a temporal interval. After filtering the signal, an unsupervised clustering is employed for finding video scene boundaries.

A common deficiency of the reviewed techniques is that they are based on a set of production rules of how the program should be composed. For instance, in home-video scenes, the shots are generally long, and their motion content is high. On the other hand, the shots are short and the visual appearance is smooth in TV-series scenes. However, these heuristics are not applicable to the different genres of videos. Another deficiency is that they ignore the context information of neighboring shots to segment scenes. Due to semantic connectivity in scenes, the features of neighboring shots provide important context information to judge scene boundaries, which can effectively decrease miss or falseness of segmentation.

In this paper we propose a novel approach for scene segmentation. The most important novelty in this paper is that we transform scene segmentation into label identification problem and employ conditional random fields (CRFs) technique to predict labels. Based on the idea of supervised learning, we can train a reasonable CRFs model that adapts readily to the different genres of videos according to training data. Moreover, in the process of model training, CRFs technique adequately utilizes the context information of neighboring shots, so it delivers significantly more accurate results than previous methods. The rest of the paper is organized as follows: Section 2 describes the proposed framework in details; Section 3 shows some experiment results and analysis; Section 4 concludes our work.

2 Scene Segmentation Algorithm

According to the observations on a great number of actual data, positional property of shots in the scene can be classified into three categories: begin shot (BS) is a start point of the new scene; end shot (ES) indicates that a scene ends in this shot; middle shot (MS) is the internal shot between BS and ES. Such structures

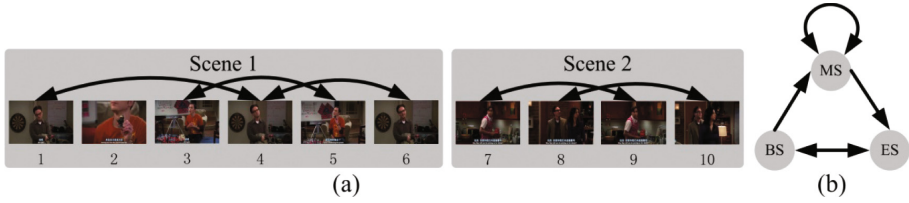


Fig. 1. (a) Repeating shot pattern between different scenes (similar shots are connected by arrows); (b) State transition graph of different types of shots in the scenes

are given in Figure 1 (a) in which BS $\{1,7\}$, MS $\{2, 3, 4, 5, 8, 9\}$ and ES $\{6, 10\}$ construct two typical scenes. In the scene sequence, repeating shot pattern of one person, a group of persons or the same setting can help identify the types of shots. BS is only similar with following shots, while the similar shots of ES only exist in the preceding ones; see shot 7 and shot 6 in Figure 1 (a) (similar shots are connected with arrows). MS connects with not only following shots but also preceding shots by visual similarity; see shot 4 in Figure 1. Stated thus, every shot in a scene has its positional property, so scene segmentation can be transformed into label identification problem. If we correctly identify all labels, we can obtain the scene boundaries between ES and BS.

To accurately identify the types of shots in a scene, the proposed method employs CRFs technique that is an effective algorithm to predict multiple variables depending on each other. CRFs technique has several advantages on scene segmentation. Firstly, based on the idea of supervised learning, CRFs technique trains a reasonable model that reveals the production rules of scene. As stated above, the shot types can be identified by production rules. It is difficult to enumerate all the production rules of the different programs, but these rules can be simulated by a CRFs model according to training data. Secondly, CRFs model integrates more statistical information comparing with graph-based method. On one hand, CRFs model estimates the state transition probabilities, when the states transform between different types of label. State transition relationship is given in Figure 1 (b), where direction of arrows indicates the only way of state transition and state transition probabilities are calculated by training data. These transition directions and probabilities simulate the retaliation of different labels in the shot sequence conducting reasonable results of label estimation. On the other hand, CRFs technique counts all priori probabilities of the three states, which promote the accuracy of label estimation. For example, if there are not effective features to identify type of a shot, it will count the most likely label from training data as the outputting label. Finally, comparing with HMM or other models, features of neighboring shots are taken into account when predict the current label. In Figure 1 (a), shot 4 is similar with shot 1 and shot 6, so the features of these shots provide context information to judge it as a MS. CRFs model integrates the above advantages, so it can accurately predict label of shots for scene segmentation.

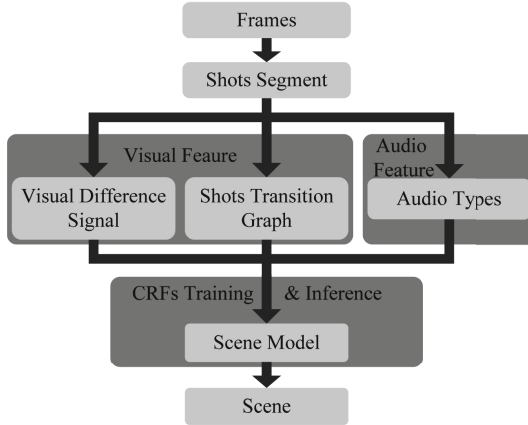


Fig. 2. Flow chart of our algorithm

In Figure 2, we summarize the main steps of our approach. The video is divided into shots using algorithm in [9]. Next, visual and audio features are extracted based on shot units, and then those features are discretized to train a CRFs model or predict the labels. In learning process, we choose a certain range of neighboring features and build a list of feature vectors that map answer tags to train a model against scene segmentation. In the predicting process, we use the model to predict labels in testing data.

2.1 Features

Visual features are directly extracted from key-frames in shot units. To reduce computational complexity, we employ a common sampling strategy to select key-frames. Assuming a sampling step is n_t , and n_s is the number of frames in one shot. When $n_s > 3n_t$ key-frames are sampled by step n_t , otherwise the first, middle and last frame are selected as the key-frames. Using this strategy, no less than three key-frames would be selected to represent each shot, which is enough to compute features. In this work, we choose three middle-level features: shot difference signal (SDS), scene transition graph (STG) and audio type (AT).

Shot Difference Signal (SDS) Feature: The first kind of features in our algorithm is SDS that depicts visual dissimilarity of scene boundary, and this signal is calculated by the graph partition function. Unlike distance between low-level features, SDS considers visual information of neighboring shots in a temporal interval, so it produces a signal with local invariance. Before calculating SDS, visual distance between each two shots must be defined according to RGB histogram metric of the key-frames. To be robust to noise, the metric in [10] is used. The two key-frames from different shots are divided into 16 blocks of the same size, as shown in Figure 3 (a). A 48-bin RGB normalized color histogram

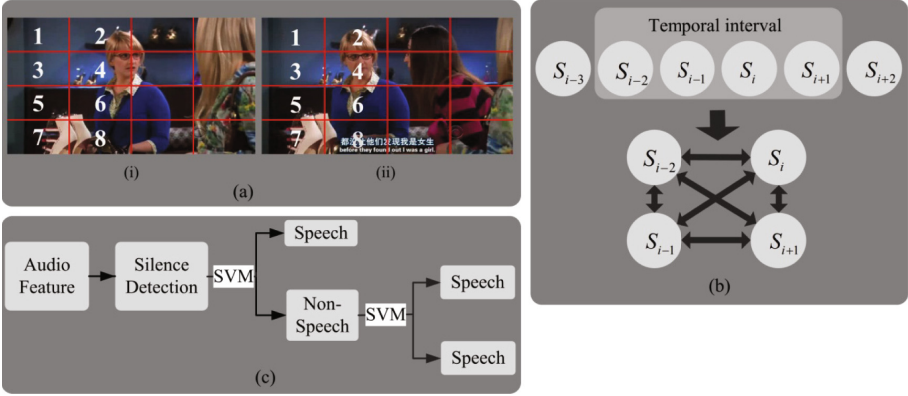


Fig. 3. (a) Eight regions of the two key frames with similar RGB histograms; (b) A typical graph-based representation of a temporal interval; (c) Flow chart of audio classify

for each region with 16 bins in each color space is extracted. Distance between corresponding blocks is calculated as follows:

$$d = 1 - \sum_{i=1}^{48} \min(H_m^i, H_n^i) \tag{1}$$

Eight regions with the largest d are discarded to reduce the effects of object motion and noise, and the metric D_k between two key frames is defined as the mean of the distances of the remaining regions, as shown in Figure 3 (a). Visual distance of shots D_s equals to the minimum D_k between two groups of key frames. The graph partition model of computing shots difference signal is explained as min-max cut [8]. All shots V in the temporal interval are partitioned into two disjoint subsets A and B , $A \cup B = V$ and $A \cap B = \phi$. Min-max cut criterion is defined as follow:

$$Mcut(A, B) = \frac{cut(A, B)}{assoc(A)} + \frac{cut(A, B)}{assoc(B)} \tag{2}$$

where $cut(A, B)$ and $assoc(A/B)$ are defined as follows:

$$cut(A, B) = \sum_{i \in A, j \in B} D_s(i, j) \tag{3}$$

$$assoc(A/B) = \sum_{i, j \in A/B} D_s(i, j) \tag{4}$$

If a $2l$ shot sequence is considered, the signal of min-max cut is calculated as follows:

$$score(i) = Mcut(A, B) = Mcut \{ \{S_{i-l}, \dots, S_{i-1}\} \{S_i, \dots, S_{i+l-1}\} \} \tag{5}$$

where integer i is an index of shot that between $[i - l, i + l - 1]$. Figure 3 (b) illustrates a typical graph-based representation of a temporal interval when $l = 2$.

The SDS must be discretized in order to input CRF++ tools [11] that are open source tools for CRF application. Since the scene boundaries are probably obtained at its local maximum and this value should be bigger than median or mean, the discrete features must reflect these characteristics. On one hand, range of the signal is divided into thirteen equal subintervals to map each value in signal sequence. On the other hand, each value is labeled three attribute: above or below median, above or below mean and local maximum or not. Therefore, SDS is transform into 4 dimensional discrete features to input into CRFs model.

Scene Transition Graph (STG) Feature: In contrast to shot difference signal depicting visual dissimilarity, STG [2] clusters similar shots for purpose of constructing connecting graph that depicts repeating shot pattern in a scene. The cut-edges of this graph are candidates of scene boundaries. To calculate this feature, shot difference D_s in SDS is employed. In clustering step, we choose a minimum spanning tree (MST) clustering in which time-constraint is easily added in clustering process. For the MST clustering, the distance matrix $A_{N \times N}$ and element $a(i, j)$ in the matrix are expressed as follow:

$$a(i, j) = \begin{cases} D_s(i, j) & \text{if } |i - j| < \sigma \\ 1 & \text{if } |i - j| \geq \sigma \end{cases} \quad (6)$$

If temporal distance between shots i and j is larger than threshold σ , they must belong to different scenes $a(i, j) = 1$, which is time-constraint in MST clustering. Object clusters can be grouped through the following steps:

1. Calculate the minimum spanning tree (MST) of matrix $A_{N \times N}$.
2. Cut the edges whose weights exceed a threshold γ in the MST forming a forest.
3. Find all the trees contained in the forest and consider each tree as a potential cluster.

STG can be constructed by backward searching in the same cluster as [2]. According to result of STG analysis, each shot is classified into two categories: boundary of STG and interior node of STG, which are discrete STG features.

Audio Type (AT) Feature: A change between scenes in the TV-program commonly accompanies a certain audio type. Therefore, AT is an effective clue that indicates a starting point of new scene. For example, there may be a silence or music appearing between different scenes. A helpful audio classification algorithm in [12] is used to classify sound types. Features are extracted from audio data across two shots during half second in each one. Then, all sounds are classified into silence, speech, music and noise by two cascaded SVMs, as shown in Figure 3 (c).

2.2 Scene Model Based on CRFs

There are many articles to introduce principle of CRFs, such as [13] [14], and due to the limited space, we will not go into these details of CRFs in this paper. To implement our algorithm, we use CRF++ tools to train the model and predict labels. We use 6 dimensional features in which 4 components are SDS, 1 component is STG and 1 component is AT. To train a CRFs model, each feature vector must map a tag to indicate the type of shot. In predicting process, we only input a feature vector list to obtain scene boundaries between ES and BS.

Let $S = \{s_i, i \in n\}$ represent n labels of shot sequence, and $X = \{x_i, i \in n\}$ is corresponding feature vector sequence. Each x_i in X represents a group of audio and visual features extracted from shot i . The goal of scene segmentation is to maximize the number of labels s_i that are correctly classified, which need to learn an independent per-position classifier that maps $X = \{x_i\} \rightarrow S = \{s_i\}$ for each shot i . The solution of CRFs to this problem is to model the conditional distribution $p(S|X)$. The probability assigned to a label sequence for a particular sequence of shots by a linear-chain CRFs is given by the equation below:

$$p(S|X) = \frac{1}{Z(X)} \exp \left(\sum_{i=1}^n \sum_{k=1}^m \lambda_k f_k (s_{i-1}, s_i, x_i) \right) \quad (7)$$

where $Z(X)$ is a normalization function:

$$Z(X) = \sum_{s_i \in S} \exp \left(\sum_{i=1}^n \sum_{k=1}^m \lambda_k f_k (s_{i-1}, s_i, x_i) \right) \quad (8)$$

function $f_k(\cdot) \in \{0, 1\}$ represents empirical function which depends on input variable. In theory, current label s_i can depend on the feature vectors of all shots, but the feature vectors of neighboring shots are only considered in practice. In formula (8), m donates range of neighboring feature vectors to predict s_i . CRF++ tools use a template to control the value of k , and the details can refer to [11]. Using $\lambda = \{\lambda_k, k \in m\}$ that is estimated in learning process, the maximum probability of the label sequence $S = \{s_i, i \in n\}$ in the condition $X = \{X_i, i \in n\}$ can be calculated, which $S = \{s_i, i \in n\}$ is desired of label sequence.

3 Experiment Results

We choose about 3 hours and 50 minutes data to train CRFs model and 6 hours and 30 minutes data to evaluate the performance of our method. Training data set and testing data set are same style but there are not overlap between them. The data sets that contain seven kinds of TV-program can verify the effectiveness of our method in scene segmentation. Table 1 summarizes the information of training and testing data set. As shown in Table 1, feature film contains a lots of conversation; action film contains car chases and gun fights; sitcom is situation

Table 1. The information of data set

genre	Testing data				Training data			
	segment	time(min)	shots	scene	segment	time(min)	shots	scene
cartoon	2	36	551	19	1	20	232	11
feature-film	2	60	477	26	1	30	258	14
action-film	2	61	1239	34	1	30	620	18
sitcom	4	84	1510	39	1	22	363	12
TV-series	2	60	1239	42	1	30	542	19
home-video	2	50	47	12	2	40	36	12
documentary	2	40	571	62	2	60	613	21
total	16	391	5634	234	9	232	2664	107

comedy; TV-series is a series of long television play; home-video is the personal or consumer video. For each video, ground-truths of the scene boundaries are obtained by a human observer in accordance with definition in work [8]. Recall, precision, and F-measure are selected following the work [8] to evaluate the performance. In addition to use CRF++ tools, we also implement our method and comparative methods by C++ language and Opencv tools.

3.1 Comparison of Different Templates

As mentioned above, the template is employed to control the range of neighboring features used to predict current label in CRF++ tools. In the first experiment, we compare the results using different templates on scene segmentation. In Figure 4, the performance of our algorithm are respectively presented by varying the template from 1 to 4. It can be observed that the algorithm yields better results with template increasing. The probable reason for this phenomenon is that large template provides more context information on predicting labels than small one. For example, repeating shots are usually separated by other shots in the same scene, as shown in Figure 1 (a). The large template is more possible to contain these features that are important clues to identify shot types.

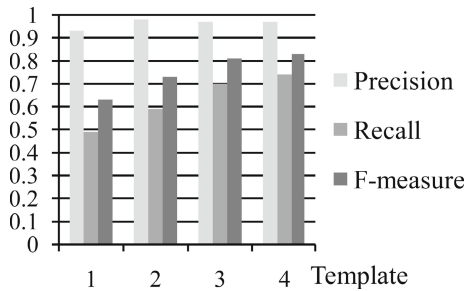


Fig. 4. Comparative results of scene segmentation using different templates in CRFs model

3.2 Comparison of Different Methods

To prove the effectiveness of our approach in scene segmentation, we have also compared with two other methods [8] and [2], as shown in Table 2. Algorithm [2] is a classical method that is compared with many works, such as [7,5,3] et al, so it can be seen as a baseline. This method groups the shots by time-constrained clustering and use grouping results for building a STG. Algorithm [8] uses graph partition model to construct a one-dimensional signal and recognizes video scene boundaries by clustering on this signal, which represents the state-of-the-art technique of scene segmentation. In comparative methods, we get a compromise between precisions and recalls listing the best F-measures in Table 2. Excluding the results of home-video, the averages of recall in three methods, 0.85, 0.81 and 0.83 respectively, are very close. By manually browsing the results in the video, we found that most right scenes were same. A possible explanation is that the features in three methods depict similar scene patterns. However, precisions of our method have obvious improvement. For CRFs model, if a pattern of a shot does not appear in training data, it tends to predict the label as the most likely type that is counted according to training data, so CRFs model can effectively control false detections and provide better precisions than other methods.

For home-video, both precisions and recalls in our method are better than other two methods. Scenes in home-video have some strong own pattern expressed by features. Algorithm [2] and [8] do not design against this pattern leading to the results not as good as other data. In our method, we train reasonable model against home-video, so the results have obvious improvement.

Results of action-film and documentary in our method are below the average. Weak efficacy of the features to depict scene patterns in action-film is responsible

Table 2. Comparative results with other methods in scene segmentation using precision, recall and F-measure

	Our method			Method in [2]			Method in [8]		
	Precision	Recall	F-measure	Precision	Recall	F-measure	Precision	Recall	F-measure
cartoon01	1	0.9	0.95	0.82	0.9	0.86	0.82	0.9	0.86
cartoon02	1	0.89	0.94	0.73	0.89	0.8	0.73	0.89	0.8
feature-film01	1	0.91	0.95	0.83	0.91	0.87	0.91	0.91	0.91
feature-film02	1	0.87	0.93	0.85	0.73	0.79	0.92	0.8	0.86
action-film01	0.63	0.71	0.67	0.79	0.88	0.83	0.83	0.88	0.86
action-film02	0.73	0.65	0.69	0.73	0.65	0.69	0.77	0.59	0.67
sitcom01	1	0.88	0.93	0.67	0.75	0.71	0.67	0.75	0.71
sitcom02	1	1	1	0.56	0.83	0.67	0.55	1	0.71
sitcom03	1	0.85	0.92	0.71	0.77	0.74	0.77	0.77	0.77
sitcom04	1	0.92	0.96	0.69	0.75	0.72	0.64	0.75	0.69
telepaly01	1	0.86	0.92	0.86	0.9	0.88	0.91	0.95	0.93
telepaly02	1	0.95	0.98	0.86	0.86	0.86	1	0.95	0.98
home-video01	1	0.8	0.89	0.5	0.4	0.44	0.5	0.4	0.44
home-video02	1	0.86	0.92	0.4	0.29	0.33	0.6	0.43	0.5
documentary01	0.77	0.74	0.75	0.85	0.81	0.83	0.81	0.78	0.79
documentary02	0.78	0.71	0.75	0.7	0.66	0.68	0.74	0.66	0.7
average	0.93	0.84	0.88	0.72	0.75	0.73	0.76	0.78	0.76

for this phenomenon. For example, few repeating shots cause the invalidation of SDS and STG. It is concluded that performance of CRFs model depends on the effectiveness of features. The ideal performance on action-film and documentary can be obtained by adding other effective features. Fortunately, CRFs model has ability to accept a large number of input features for prediction.

3.3 Accuracy of Scene Boundaries

In the last experiment, we compare accuracy with different methods. In scene segmentation, a reasonable boundary error tolerance is no more than two shots. Average length of the shots near two seconds, so average of boundary error is no more than five seconds, which is acceptable according to audience experience. However, if a method can provide more accurate results, the feeling of audience in browsing video will be better. In the experiment, we find that CRFs model tends to miss the boundaries, but accuracy of right labels is better than other methods, as shown in figure 5. When error tolerance becomes more rigorous, there are slight F-measure declines using CRFs model. The reason of this phenomenon is that CRFs model chooses the global optimum solution during inference.

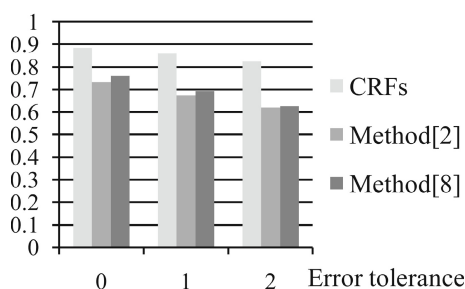


Fig. 5. Comparative results (using F-measure) with different error tolerances in scene segmentation

4 Conclusion

In this paper a novel scene segmentation method, making use of CRFs technique, is presented. As the contribution of this method, algorithms is developed for scene segmentation by transforming it into label identification problem, and CRFs modal is exploited to identify labels sequence by three kinds of middle-level features that depict the properties of the shot. Since the context information of neighboring shots is taken into account, CRFs model delivers accurate results on scene segmentation. In experiment, our method is successfully validated on various types of video and the encouraging experimental results demonstrate its effectiveness to scene segmentation.

References

1. Petersohn, C.: Logical unit and scene detection: a comparative survey. In: *Multimedia Content Access: Algorithms and Systems II*, vol. 6820, pp. 2–17 (2008)
2. Yeung, M., Yeo, B.L., Liu, B.: Segmentation of video by clustering and graph analysis. *Computer Vision and Image Understanding* 71, 94–109 (1998)
3. Chong-Wah, N., Yu-Fei, M., Hong-Jiang, Z.: Video summarization and scene detection by graph modeling. *IEEE Transactions on Circuits and Systems for Video Technology* 15, 296–305 (2005)
4. Jianbo, S., Malik, J.: Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 888–905 (2000)
5. Rasheed, Z., Shah, M.: Detection and representation of scenes in videos. *IEEE Transactions on Multimedia* 7, 1097–1105 (2005)
6. Yun, Z., Shah, M.: Video scene segmentation using markov chain monte carlo. *IEEE Transactions on Multimedia* 8, 686–697 (2006)
7. Chasanis, V.T., Likas, A.C., Galatsanos, N.P.: Scene detection in videos using shot clustering and sequence alignment. *IEEE Transactions on Multimedia* 11, 89–100 (2009)
8. Sakarya, U., Telatar, Z.: Video scene detection using graph-based representations. *Signal Processing: Image Communication* 25, 774–783 (2010)
9. Jinhui, Y., Huiyi, W., Lan, X., Wujie, Z., Jianmin, L., Fuzong, L., Bo, Z.: A formal study of shot boundary detection. *IEEE Transactions on Circuits and Systems for Video Technology* 17, 168–186 (2007)
10. Xinbo, G., Xiaou, T.: Unsupervised video-shot segmentation and model-free anchorperson detection for news video story parsing. *IEEE Transactions on Circuits and Systems for Video Technology* 12, 765–776 (2002)
11. Kudo, T.: Crf++: Yet another crf toolkit (2005)
12. Li, Y., Dorai, C.: Svm-based audio classification for instructional video analysis. In: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*, vol. 5, pp. 897–900 (2004)
13. Sutton, C., McCallum, A.: *An Introduction to Conditional Random Fields*. ArXiv e-prints (2010)
14. Klinger, R., Tomanek, K., Klinger, R.: *Classical probabilistic models and conditional random fields* (2007)

Improving Preservation and Access Processes of Audiovisual Media by Content-Based Quality Assessment

Peter Schallauer, Hannes Fassold, Albert Hofmann, Werner Bailer,
and Stefanie Wechtitsch

JOANNEUM RESEARCH Forschungsgesellschaft mbH – DIGITAL,
Steyrergasse 17, 8010 Graz, Austria

{Peter.Schallauer,Hannes.Fassold,Albert.Hofmann,Werner.Bailer,
Stefanie.Wechtitsch}@joanneum.at

Abstract. Quality assessment of audiovisual files is an important tool in many steps of the preservation workflow, as well as for use and access of archive material. Today mainly technical properties of the files can be checked, e.g. file integrity or standards compliance of file wrappers and encoded streams. Checking the audiovisual quality manually results in extremely high labor costs. In this work we present a semi-automatic quality assessment approach that combines the efficiency of fully automatic detection with the interpretation capability of humans to provide verified high quality assessment results. We also address the issue of interoperable metadata for quality assurance, discussing the state of the art and the gaps, and propose a framework for describing visual quality analysis results, which fills one of these gaps.

Keywords: multimedia preservation, visual quality analysis, preservation metadata.

1 Introduction

Quality control for audiovisual (AV) files is an important tool in several steps of the preservation and access processes of audiovisual archives, including ingest, restoration and delivery. In these processes AV files need to be checked on different levels. On the file level data integrity is checked by means of fixity information, e.g. with checksums or hashes implemented within storage, content or preservation management systems. On the file wrapper (e.g. MXF) and on the stream encoding (e.g. MPEG-4 AVC/H.264 or JPEG2000) level standards compliance is checked by means of available industry tools. On the content level only a small set of tools for spotting visual or audio distortions is available. Today manual checking of the visual and audio quality results in extremely high labor costs or, if these costs cannot be afforded, in not assuring the content quality. The transition to file based production and preservation environments enables automation of quality assurance tasks with the goals of reducing quality assurance costs, increasing the quality of the content produced and assuring quality of content ingested in and re-used out of a digital archive.

The rest of this paper is organized as follows. Section 2 presents use cases for content based visual QA in preservation processes. Section 3 provides an overview on our novel tools for automatic quality detection and Section 4 discusses a user interface for efficient interactive quality verification. In Section 5 we address the issue of interoperable metadata for quality assurance, discussing the state of the art and the gaps, and propose a framework for describing visual quality analysis results, which fills one of these gaps. Section 6 concludes the paper.

2 Use Cases for Content Based QA in Preservation and Access

Content based quality assessment can be beneficial in various digital video or movie archive related use cases. During archive content ingest or migration, it is useful for monitoring the film scanning process, e.g. for white and black points, instability, out of focus, flicker, etc., for the monitoring of video player problems like head clogs, drop-outs, video breakups, off-lock situations and for checking the encoding or transcoding for blocking and blurriness artifacts. Furthermore, for the process of archive content selection, access and usage it can be a valuable tool for several tasks like selecting the ‘best quality copy’ in case the content is available in more than one copy, for selecting a video or movie with minimum quality for a certain usage (e.g., is the actual resolution of the content suitable for standard definition to high definition up-conversion/broadcast/Blu-ray production) and for selecting a movie where additional post processing costs can be avoided (e.g., a movie with low film grain noise/flickering/image instability), thereby avoiding restoration costs. Finally, content based quality estimation can be used in restoration planning to estimate the cost and time of restoration (based on the amount of dust, noise, flicker and image instability present in the un-restored content) and to select the most appropriate restoration tools/systems.

Although humans are able to provide very reliable quality assessments, in practice this approach is too time-consuming and therefore extremely costly. Alternatively, fully automatic QA approaches can be implemented very cost efficiently, but cannot provide the same functionality and reliability as human judgments. We thus aim at an approach combining the benefits of both worlds, i.e. the cost efficiency of automatic tools with the interpretation capability of humans. Fig. 1 illustrates this semi-automatic approach where audiovisual files are first analyzed fully automatically and then these analysis results are verified interactively by humans, providing a final quality report.

An analysis profile defines the type and parameters of automatic impairment detectors to be applied for a certain QA task, e.g. film scanning QA, video tape migration QA and restoration preparation QA.

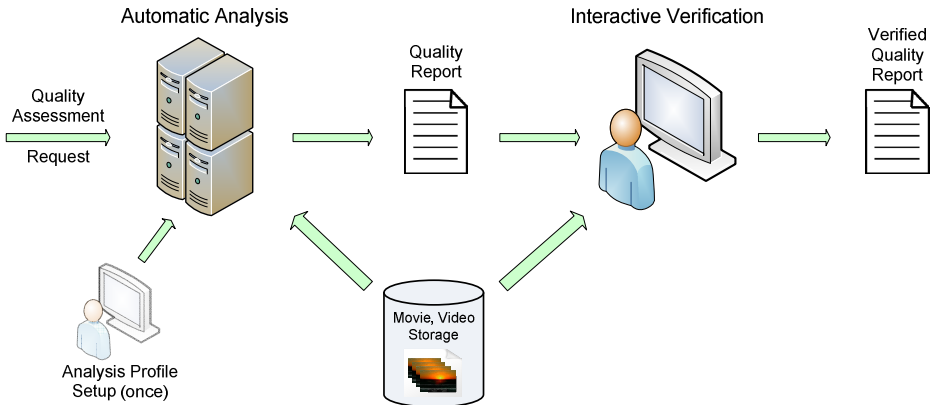


Fig. 1. Semi-automatic quality assessment workflow consisting of fully automatic quality analysis and interactive quality verification

3 Automatic Content Based Detectors

Video quality analysis methods can be categorized by the amount of information available from the original undistorted video. Full-reference methods need access to both, the current copy and the original undistorted video (the reference). Reduced-reference QA methods work with only partial information of the original video. Non-reference methods require only the current copy available. In the application areas preservation and restoration the original undistorted video is practically unavailable in almost any case; therefore non-reference QA methods allow the widest range of application. In the following a set of state of the art detectors based on non-reference methods is described.

Electronic noise (of analogue or digital source) or film grain noise is present to a different degree in any video or movie content. Also modern digital video cameras create a significant amount of noise when shooting e.g. in low-light conditions. The noise/film grain detector supports noise level estimation for different types of noise, from very fine electronic noise, over different kinds of digital sensor noise up to very coarse film grain noise. Interlaced as well as progressive sampled video/movie is supported. The signal dependency of noise (different strength of noise in different luminance and chrominance channels) is also estimated. The detector is robust against a wide range of image degradations including flicker and image instability. The noise/grain value is estimated for every n^{th} frame statistically from measurement values of a temporal sliding window centered at the frame. A graphics board with CUDA¹ support is needed for the computationally expensive local motion estimation task, allowing the detector to operate in real-time for standard definition (SD) material. The detector can be used for determining whether noise restoration/

¹ Compute Unified Device Architecture (CUDA) is a parallel computing architecture developed by NVIDIA for general purpose processing on graphics cards.

reduction is required, e.g. in post-production, before archive content re-use or before play-out. It can also be used to monitor the film scanning process in regard to noise produced by the scanner.

The video breakup detector described in [1] and [2] detects temporal segments in the video containing major image disruptions, for example caused by head clogging, assemble edits, lost lock, recorded serious digital error corrections, severe TBC hits and damaged tapes. A typical video breakup defect is shown in Figure 2 at the lower-right. Although the detector primarily targets analogue defects (showing horizontal line distortions), also severe digital errors (typically exhibiting blocking defects) are detected. The output of the video breakup detector is a list of temporal segments where video breakups occur, and a severity value for each segment which gives an indication of how severe the video breakup defect is. Due to CUDA support for key components like the motion estimation, the detector is able to operate in real-time for SD material.

The sharpness detector measures the actual sharpness of the content, relative to the nominal video/movie resolution. For the estimation of the image sharpness, the image is divided into blocks and a sharpness value is calculated for each block from the edge widths of the horizontal and/or vertical edges within the block. From the block sharpness values, the image sharpness is then calculated with robust statistical methods. Additionally, the detector can take the human perception of sharpness into account by focusing only on the most significant edges in the image. The detector is robust against common degradations appearing in video and film like noise, flicker and instability. The sharpness value is calculated for every n^{th} frame in the video. It can be used for example to detect if content has been up-scaled from standard definition to high definition video, or to monitor the film scanning process for out of focus quality control.

Freeze frames occur, when no valid content data for the current frame can be retrieved due to various reasons. In this case, most video player or transmission devices deliver the previous frame (or field) instead, leading to several consecutive frames with identical content in the video stream. The freeze frame detector described in [3] detects temporal segments where freeze frame defects occur. The detector is able to differentiate normal static image content (e.g. titles and caption) from an actual freeze frame defect and is robust against noise which can be superimposed on frozen frames. The detector operates significantly faster than real-time for SD video material.

Dust, dirt and blotches are very frequent defects in archived film. They appear as bright or dark spots of irregular shape and have in common that a single defect of this category occurs usually only in one frame. Utilizing this characteristic, the single-frame defect (SFD) detector described in [4] measures the amount of dust, dirt and blotches in a frame as percentage of the image area occluded by these defects. The SFD defect amount is calculated on a temporally sub-sampled video, e.g. for every 20th frame. The SFD detector operates in real-time for material in SD resolution.

4 Efficient Manual Quality Verification

Efficient visualization and verification of impairment analysis results supports an operator to get a quick overview of the condition of the material and to allow for manual corrections and final quality judgment by the operator.

In the following we describe the user interface shown in Fig. 2, which is composed of these four main parts: Global timeline views (1) show the occurrence of defect events for the full temporal range of the video. A global timeline view also shows the shot structure and the temporal zoom period for the timeline views in part (3). For efficient verification, a defect list component (2) shows defect events and their properties. Timeline views showing a zoomed temporal resolution providing a level of temporal detail that can be freely adjusted are shown in part (3) of the user interface. A video player (4) with frame accurate positioning support and audio playback is also provided.

The video player is the central component of the user interface. All other components synchronize with it. The video player can be positioned on an extra monitor for full resolution playback. The other components like the event list component and the timeline views can also be displayed on a second monitor.

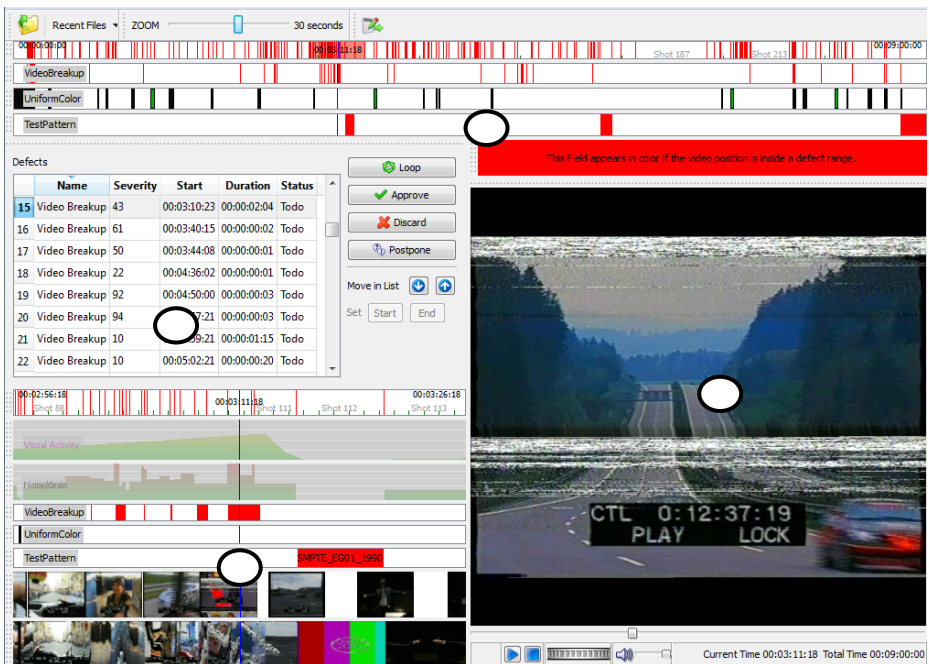


Fig. 2. User Interface for efficient interactive verification of automatic detections

All components provide additional navigation functionalities. The key frame and stripe image timeline views shown in the bottom of part (3) provide a quick visual overview of the video content. Key frames and stripe images are aligned on the timeline according to their respective time points. Navigation is possible by clicking on the timeline, or by moving the scroll wheel for frame accurate positioning.

Timeline views showing impairment detection results either visualize continuous quality measures in form of line or bar charts like the visual activity and the noise/grain level within specific time ranges. Detections having an event-like character are also visualized on timeline views by indicating the temporal segment of the detection. These are for example video breakups, uniform color and test pattern segments. The different views appear both over the full video range in part (1) and for the selected zoom period in part (3). For uniform color detections the respective segments are additionally filled with the color detected.

The time an operator can spend to verify automatic analysis results is typically limited and it may be the case that not all defect detections can be manually verified. So the time the operator has available should be utilized best. For this it is very useful to be able to handle the most relevant detections first. To support this, the detections listed in the defect list view can be sorted by all columns. So when sorting by severity an operator can efficiently verify the most relevant detections first. A detection can either be approved, discarded, or postponed for later verification by the operator. After such a manual verification the next detection in the list not verified yet will be selected. This verification process is supported by a special mode where the video will play in a loop around the currently selected detection including a configurable pre roll and post roll time.

5 Interoperable Metadata for Quality Assurance

Interoperable metadata is a key prerequisite for long-term preservation and quality assurance of audiovisual content. For preservation purposes, the following two types of metadata are most crucial: Structural metadata, i.e., metadata that is needed to correctly interpret the stored essence (header structures of containers, technical metadata about the type of encoding, etc.), and preservation metadata, which includes information about the fixity of the object (i.e., properties that allow checking the integrity and quality of the essence), as well as a documentation of the preservation actions applied (e.g., devices/tools used and their parameters).

Here we focus on metadata for describing processing applied in the preservation (e.g., digitization) workflow, which might provide valuable information about content quality based on measurements from tape recorders, etc., and on the representation of information coming from content quality checking tools as described in the previous sections. The description of the quality of audiovisual content and the defects it might contain is part of the content's preservation metadata. Thus, the metadata cannot be dependent on the specific tools used for quality analysis, but must be understood by all the different preservation and restoration tools in the workflow. In addition, preservation metadata is only useful if it can still be interpreted many years after its creation. These requirements call for a standardized approach for representing quality metadata.

The quality description shall allow getting an overview of the condition of the audiovisual material. It shall thus be a compact description and contain details only if absolutely necessary. The description is mainly produced by automatic tools, only validated by the operator, and it shall also be possible to process the description automatically. Therefore, the time point or range for which a description is valid must be specified, quality has to be quantified numerically or by sets of defined terms, defects need to be unambiguously identifiable, and optionally, properties of defects may be further described numerically or by sets of defined terms.

As the descriptions support the user in getting a quick overview of the materials condition, they shall be defined in a way that they are easy to visualize. Especially quality measures and defect descriptors that represent a larger time range shall allow condensed visualization over time. Quantitative descriptions of impairments shall correspond to the perceived severity of the defect.

5.1 State of the Art

There are several gaps in existing metadata standards for this type of information. MPEG-7 [5] is a standard for the description of multimedia content, including structuring the content as well as describing a number of low-, mid- and high-level features for each of the segments in the structure. In MPEG-7 some impairment descriptors and description schemes have already been standardized. The MediaQuality descriptor contains (i) a quality rating, expressed as a floating point value, (ii) a rating source (iii) a list of perceptible defects, discriminated into visual and audio defects, each of them being a term in a classification scheme. Classification schemes are MPEG-7 description schemes for defining hierarchies of controlled vocabulary. It is however not possible to describe the defect in more detail or its exact (spatio-) temporal location. The AudioSignalQualityDS [6] can be added to each audio segment and contains some segment-based audio parameters and a list of error events. Each of these error events is described by the error class (a reference to a term in a classification scheme), time stamp and channel number, detection method (manual, automatic), relevance, status and optional text annotation.

The SMPTE metadata dictionary [7] contains a long list of properties, including many describing tools and their settings in production, but lacks comparable properties for preservation. First steps to document the knowledge involved in preservation processes are simple databases of format properties on obsolescence (e.g., PRONOM [8], JHOVE [9]), but they do not sufficiently cover the audiovisual domain. The PrestoPRIME project² has started working on a registry including information about the obsolescence of audiovisual formats, which also considers the issues of container formats with different encodings inside. Very recently, the MPEG Multimedia Preservation Description Information (MPDI) group has identified these and a number of related issues in their requirements document.

5.2 MPEG-7 Extension for Visual Defect and Quality Description

Based on an analysis of the state of the art and the requirements defined above it becomes clear that MPEG-7 is a suitable standard to serve as a basis for the

² <http://www.prestoprime.eu>

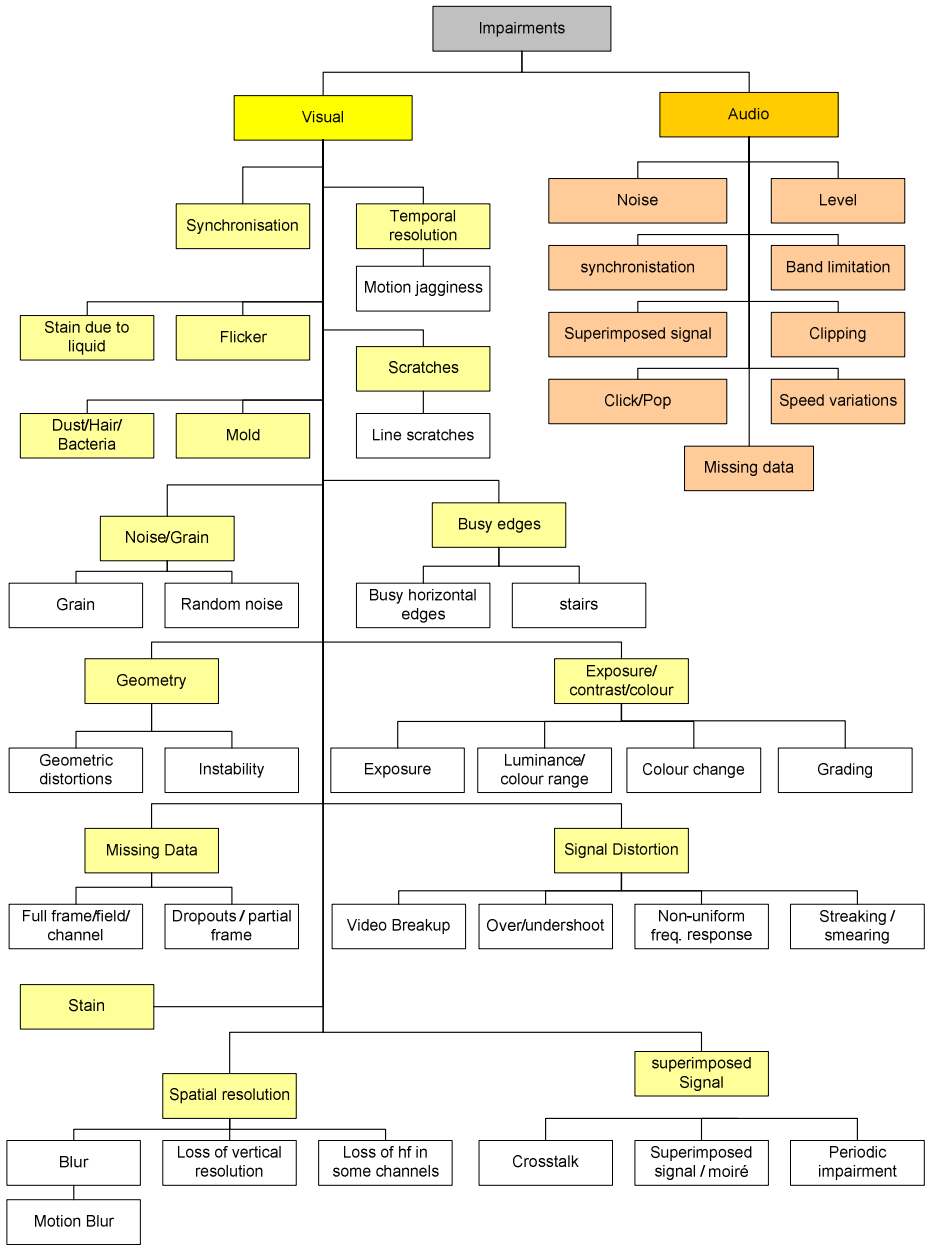


Fig. 3. Top-level classes of the impairment classification scheme

description of visual impairments. MPEG-7 allows to structure descriptions on different levels of granularity and already offers some tools for quality description, especially in the audio domain. Thus we have proposed an extension to MPEG-7 for

describing visual impairments, similar to that for audio quality and defects defined in [6], a set of detailed descriptors for specific quality measures and defect descriptors and an extended classification scheme for visual and audio impairments.

The extension is based on the MPEG-7 Audiovisual Description Profile [10] which has been proposed for detailed description of audiovisual content in production and archiving. The MPEG-7 extension for defect and quality description is available at [11].

There is a generic visual descriptor for defects which specifies general properties and references in a classification scheme. This is the minimum description of a defect, specifying its type and the segment of its occurrence. In addition, specific descriptors for a number of defects and quality measures have been defined, which allow to describe their respective properties.

Starting from the Brava broadcast archive programme impairments dictionary [12], we have defined a comprehensive impairment classification scheme that provides for hierarchical organization and multilingual description of defects. The main organization criteria of the classification scheme are the visible and audible effects of defects. The top levels of the visual and audio defects in the classification scheme are visualized in Fig. 3. This classification scheme of impairments is only a starting point, and more work on standardization of taxonomies of impairments, related devices, carriers and encoding formats is required.

6 Conclusion

Content based quality assessment is a beneficial tool in preservation of audiovisual content, covering use cases such as archive content ingest and migration, archive content selection, access and usage and restoration planning. Automation of QA is essential to reduce costs; a semi-automatic QA approach optimally combines the efficiency of fully automatic detection with the interpretation capability of humans to provide verified high quality results. A set of tools for automatic content-based detection of common video/movie defects like noise & film grain, test pattern, video breakup, sharpness, freeze frame and dust & dirt was presented, operating in real-time for content in SD resolution. In order to verify results, automatic detections need to be checked by human operators. A tool was presented providing various timeline views giving a quick overview of the condition of the material and supporting an operator in efficient verification and final quality judgment. In order to ensure that preservation metadata, including process related information and quality analysis results can be exchanged with tools throughout the workflow and remain long-term understandable, a description based on the MPEG-7 metadata standard has been proposed. We have also proposed a taxonomy for impairments and discussed related open issues for standardization. Our future work will focus on the development of automatic quality detection and interactive verification functions needed for the large number of additional impairments occurring in movie and video production, preservation, migration and re-use scenarios.

Acknowledgements. The authors would like to thank Harald Stiegler, Martin Winter and Georg Thallinger as well as several other colleagues at JOANNEUM RESEARCH, who contributed valuable input to the work. This work has been funded partially under the 7th Framework Programme of the European Union within the ICT projects "PrestoPRIME" (ICT FP7 231161) and "DAVID" (ICT FP7 600827) and under the FIT-IT Programme of the Austrian Federal Ministry for Transport, Innovation and Technology within the project "vdQA".

References

1. Winter, M., Schallauer, P., Hofmann, A., Fassold, H.: Efficient video breakup detection and verification. In: Workshop on Automated Information Extraction in Media Production (2010)
2. Rosner, J., Fassold, H., Winter, M., Schallauer, P.: Real-time video breakup detection for multiple HD video streams on a single GPU. In: SPIE Real-time Image and Video Processing (2012)
3. Schallauer, P., Fassold, H., Winter, M., Bailer, W.: Automatic freeze frame detection for video preservation. In: IEEE International Conference on Image Processing (2009)
4. Schallauer, P., Bailer, W., Mörzinger, R., Fürntratt, H., Thallinger, G.: Automatic quality analysis for film and video restoration. In: IEEE International Conference on Image Processing (2007)
5. ISO/IEC 15938:2001. Information Technology - Multimedia Content Description Interface
6. ISO/IEC 15938-4:2002/AMD1:2004. Information Technology - Multimedia Content Description Interface, Part 4: Audio, Amendment 1
7. Metadata Dictionary Registry of Metadata Element Descriptions. SMPTE RP210.11 (2008)
8. PRONOM, <http://www.nationalarchives.gov.uk/PRONOM/Default.aspx>
9. JHOVE - JSTOR/Harvard Object Validation Environment, <http://hul.harvard.edu/jhove/http://hul.harvard.edu/jhove/>
10. ISO/IEC 15938-9:2005/AMD1:2012. Information technology - Multimedia content description interface - Part 9: Profiles and levels, Amendment 1: Extensions to profiles and levels
11. Bailer, W., Schallauer, P., Noiré, J.-E., Maziewski, P., Fassold, H., Winter, M.: Audiovisual Defect and Quality Description (v1.6), Technical report (2011), <http://mpeg7.joanneum.at>
12. Brava, The Brava broadcast archive programme impairments dictionary (2002), http://brava.ina.fr/brava_public_impairments_list.en.html

Distribution-Aware Locality Sensitive Hashing

Lei Zhang^{1,2}, Yongdong Zhang¹, Dongming Zhang¹, and Qi Tian³

¹ Advanced Computing Research Laboratory, Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology,

Chinese Academy of Sciences, Beijing, China

² University of Chinese Academy of Sciences, Beijing, China

³ University of Texas at San Antonio, San Antonio, USA

{zhanglei09, zhyd, dmzhang}@ict.ac.cn, qitian@cs.utsa.edu

Abstract. Locality Sensitive Hashing (LSH) has been popularly used in content-based search systems. There exist two main categories of LSH methods: one is to index the original data in an effective way to accelerate search process; the other one is to embed the high-dimensional data into hamming space and perform bit-wise operations to search similar objects. In this paper, we propose a new LSH scheme, called Distribution-Aware LSH (DALSH), to address the problem of lacking adaptation to real data, which is the intrinsic limitation of most LSH methods belong to the former category. In DALSH, a given dataset is embedded into a low-dimensional space with projection vectors learned from data, followed by deriving hash functions from the distribution of the dimension-reduced data. We also present a multi-probe strategy to improve the query performance. Experimental comparisons with the state-of-the-art LSH methods on two high-dimensional datasets demonstrate the efficacy of DALSH.

Keywords: Similarity search, Approximate nearest neighbor search, Locality sensitive hashing, Distribution-aware hashing.

1 Introduction

High-dimensional similarity search is a fundamental problem in many content-based search systems. To solve this problem efficiently, many index methods have been proposed, such as KD-tree [1] and R-tree [2]. These methods work well for low-dimensional data, whereas do not scale well with dimensionality because of “the curse of dimensionality” [3]. When the dimensionality exceeds 10, these methods are even inferior to linear-scan [4]. Fortunately, it is sufficient to find Approximate Nearest Neighbors (ANN) in many applications [5]. Many methods for ANN search are developed and among these methods, Locality Sensitive Hashing (LSH) [6, 7] and its variants [8, 9] are well-known for their efficacy. LSH was first proposed by Indyk et al. in [6]. It has been extended to a variety of similarity metrics beyond Euclidean distance [7, 20], including hamming distance [5], p -norm distance [7] and kernel similarity [10, 21]. In general, there exist two main categories of LSH methods: one is to index the original data in an effective way to accelerate the search process (we call “original” LSH); the other one is to embed the high-dimensional data into hamming space and perform bit-wise operations to search similar objects (called binary LSH).

A famous “original” LSH method is the Euclidean LSH (E2LSH) [7], which uses linear projections to partition a given dataset into several subsets. Some variations [8, 12] based on multi-probe strategy have been brought forward to reduce the storage space. In [13], LSH Forest is proposed to eliminating the different data-dependent parameters for which LSH must be constantly hand-tuned. Several other hash methods have been proposed to improve the query performance, such as Leech lattices [14] and E8 lattices [15]. In lattice-based hashing, any two points in the same lattice are separated by a bound distance which only depends on the lattice definition. The lattice-based hash functions exploit the vectorial structure of Euclidean space, thus outperform E2LSH. In [16], Paulevé et al. use the k-means clustering to partition a dataset and propose k-means LSH (KLSH). A given dataset is first clustered into several subsets, and then for each subset, its cluster center is used as the identifier. Given a query, only the subset with the nearest center to the query is searched. It is argued that KLSH is the most effective algorithm of LSH [16]. To our knowledge, it is the most effective LSH method which belongs to the “original” LSH methods.

The other kind of LSH methods, binary LSH, and more generally, binary hashing, is becoming increasingly popular for efficient nearest neighbor search. Given a dataset, binary hashing generates binary code for each object and approximates the distance or similarity of two objects by the hamming distance between their binary codes. To generate compact binary code, many algorithms have been proposed. In semantic hashing [17], the Restricted Boltzmann Machine is used to map similar objects to similar codes. Spectral hashing [18] utilizes the simple analytical eigenfunction solution of 1D Laplacians as binary hash function. Shift-Invariant Kernel Hashing, a distribution-free method based on the random features mapping for shift-invariant kernels, is proposed in [19] and has a comparable performance to that of Spectral Hashing. Moreover, to exploit the spectral properties of the data affinity (e.g. pairwise similarity) to generate better binary codes, many other algorithms, such as Semi-supervised Sequential Projection Hashing [9], Anchor Graph Hashing [11] and Kernel-Based Supervised Hashing [21], have been developed and give commendable search performance. These methods are generally superior to LSH because LSH is simple, always data-independent and has no learning algorithm to reveal the underlying information of dataset. Inspired by these research works, we introduce a supervised learning algorithm to “original” LSH methods to address the problem of lacking adaptation to real data within these methods.

Hash functions of LSH determine the way a dataset indexed, hence, they are the core elements that affect the query performance. In E2LSH, the hash functions are quantized linear projections with randomly selected projection vectors. Since each projection vector is randomly selected, E2LSH partitions a given dataset without taking account of the dataset information. Therefore, the final dataset partition is not optimal and this brings some limitations. A main limitation is that multi hash tables are needed to guarantee the query accuracy. In lattice-based hashing [14, 15], the hash functions only exploit the vectorial structure of Euclidean space, whereas the hidden information of the dataset is still discarded. To make the constructed index data-adaptive, Paulevé et al. [16] have proposed k-means based LSH (KLSH). In KLSH, a given dataset is clustered into k subsets, then for each subset, the cluster center is

used as the identifier. Given a query q , only the cluster with the nearest center to q is searched. To reduce the complexity of the clustering stage, the hierarchical k-means (HKM) [16] is used. This method consists of a series of k-means with a relatively small k and produces a balanced tree structure as follows: a dataset is first clustered into k (branching factor $b_f = k$) subsets, then each subset is also clustered into k sub-subsets using k-means. This process proceeds recursively until obtaining a pre-defined tree height h_t .

Since the k-means clustering is performed in the original data space and it minimizes the inner class distances, the final dataset partition adapts to the data distribution well. It is argued that KLSH is the most effective algorithm of LSH [16]. However, k -means clustering only ensures to find local minimum, especially in high-dimensional space, hence the dataset partition is non-optimal when the dimensionality is high. Moreover, since a high-dimensional vector is used as the identifier of each subset, the memory occupation of KLSH is relatively larger than that of E2LSH.

In this paper, we propose a new LSH scheme called Distribution-Aware LSH (DALSH). It alleviates the problem of lacking adaptation to real data, which is the intrinsic limitation of most existing “original” LSH methods. Linear projection, with projection vectors learned by a supervised learning algorithm, is used to reduce the dimension of the dataset to index. Then, the hash functions are derived from the distribution of the dimension-reduced data. Experimental results on two public high-dimensional datasets demonstrate the efficacy of the proposed method as compared with the state-of-the-art “original” LSH methods, Kmeans LSH. The rest of this paper is organized as follows. In Section 2, we present our DALSH algorithm and give the multi-probe strategy to improve the query performance. Section 3 describes our experiments and Section 4 concludes this paper.

2 Distribution-Aware Hashing

2.1 Supervised Projection Learning

The hash functions of DALSH are derived from the distribution of the dataset, hence, the dataset distribution is essential. There are several novel methods that can construct predictive models for high-dimensional data with high accuracy. However, most of them are computationally expensive. To simplify the process of distribution modeling, we first use linear projection to embed the original data into a low-dimensional space. The projection vectors are learned from the data with a supervised learning algorithm.

Given a dataset $\mathbf{X} = [x_1, x_2, \dots, x_n] \in R^{d \times n}$ with a fraction of points associated with two categories of relationships: \mathcal{N} and \mathcal{C} . A tuple $(x_i, x_j) \in \mathcal{N}$ is denoted as a neighbor pair, while $(x_i, x_j) \in \mathcal{C}$ is denoted as a non-neighbor pair. Suppose there are l ($l < n$) points, each of which is associated with at least one of these two categories. The matrix formed by these l data points is denoted as $\mathbf{X}_l \in R^{d \times l}$. For a projection vector ω , we want the difference between the projections of a neighbor pair on it small while the difference between the projections of a non-neighbor pair large. As a result, the empirical accuracy of ω can be measured as follows:

$$J(\omega) = \frac{1}{2} \left(\sum_{(x_i, x_j) \in \mathcal{N}} |\omega^T(x_i - x_j)|^2 \text{sim}(x_i, x_j) - \sum_{(x_i, x_j) \in \mathcal{C}} |\omega^T(x_i - x_j)|^2 \text{sim}(x_i, x_j) \right). \quad (1)$$

where $\text{sim}(x_i, x_j) = \exp \left\{ -\|x_i - x_j\|_2^2 / \sigma^2 \right\}$ [18]. Minimizing $J(\omega)$ would make the difference between projections of a neighbor pair small; the difference between projections of a non-neighbor pair large. With simple algebra, (1) can be rewritten as:

$$J(\omega) = \omega^T \mathbf{X}_l \mathbf{L} \mathbf{X}_l^T \omega. \quad (2)$$

where $\mathbf{L} = \mathbf{D} - \mathbf{S}$, $\mathbf{S} \in R^{l \times l}$ is a label matrix defined in (3) and $\mathbf{D} = \text{diag}\{\mathbf{S}\mathbf{1}\}$.

$$\mathbf{S}_{ij} = \begin{cases} \text{sim}(x_i, x_j): (x_i, x_j) \in \mathcal{N} \\ -\text{sim}(x_i, x_j): (x_i, x_j) \in \mathcal{C} \\ 0: \text{otherwise} \end{cases}. \quad (3)$$

Let $\mathbf{W} = [\omega_1, \omega_2, \dots, \omega_K] \in R^{d \times K}$, where $\omega_1, \omega_2, \dots, \omega_K$ are projection vectors. These K projection vectors are solutions of the following optimization problem:

$$\mathbf{W} = \arg \min_{\omega_k \in R^d} \sum_k \omega_k^T \mathbf{X}_l \mathbf{L} \mathbf{X}_l^T \omega_k = \arg \min_{\mathbf{W} \in R^{d \times K}} \text{tr}\{\mathbf{W}^T \mathbf{X}_l \mathbf{L} \mathbf{X}_l^T \mathbf{W}\}. \quad (4)$$

(4) can be easily solved if the orthogonality constraints are imposed on ω_k (i.e., $\mathbf{W}^T \mathbf{W} = \mathbf{I}$). The solutions are simply the eigenvectors with the top K smallest eigenvalues of $\mathbf{X}_l \mathbf{L} \mathbf{X}_l^T$. However, the orthogonality constraint is not always reasonable and relaxing this constraints may lead to a better result [9], thus, we add a perturbation matrix $\epsilon \in R^{d \times K}$, $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_\epsilon)$ and set $\mathbf{W} = \mathbf{W}_{opt} + \epsilon$.

2.2 Distribution-Aware Hashing Function

The hash value is obtained by partitioning the data space into several subspaces with several hyperplanes perpendicular to ω_k . In this way, each data point belongs to a unique subspace and the identifier of the subspace is used as the hash value. In Section 2.1, only the labeled data are used to learn ω_k , this may lead to overfitting. To alleviate this potential problem, the information entropy of hash function, which utilizes both labeled and unlabeled data, is used as the regularizer. Denote the hash function associated with ω_k as $h_k(x)$, the entropy of $h_k(x)$ is:

$$H(h_k(x)) = \sum_z -\text{Pr}(h_k(x) = z) \log \text{Pr}(h_k(x) = z). \quad (5)$$

To maximize $H(h_k(x))$, $\Pr(h_k(x) = z)$ should be identical for all hash values z . Denote CDF_k the cumulative distribution function of projections on ω_k , we have:

$$h_k(x) = \lfloor L_k \cdot CDF_k(\omega_k^T x) \rfloor . \tag{6}$$

Each $h_k \in H: R^d \rightarrow Z$ maps a $x \in R^d$ to an integer $z \in \{0, \dots, L_k - 1\}$. It maps x to a line, and then the line is divided into L_k slots. The hash value of DALSH is a concatenation of K hash values: $g(x) = (h_1(x), h_2(x), \dots, h_K(x))^T$.

It has been reported in the literature that the distribution of projections of high-dimensional data on a randomly selected direction follows a Gaussian distribution [22]. Therefore, we use the Gaussian mixture model (GMM) with EM algorithm to estimate the distribution. The probability density function (PDF) of projections on ω_k is a mixture of Gaussian distributions: $PDF_k = \sum_{i=1}^{g_k} \pi_{ki} \mathcal{N}(\mu_{ki}, \sigma_{ki})$, where g_k is the number of Gaussian components in PDF_k .

There is a hidden problem of this hashing strategy. Since the length of slot is relatively small in dense areas, two points p, q with distance $r = \|p - q\|_2$ in a dense area are less likely to map to the same slot than two other points with the same r but in a sparse area. However, the experimental results indicate that this problem does not affect the query performance greatly. Moreover, the multi-probe strategy proposed in Section 2.3 can alleviate this hidden problem efficiently.

2.3 Multi-probe Strategy

A hash *perturbation vector* is defined as: $\Delta = (\delta_1, \delta_2, \dots, \delta_K)^T$. Given a query q , not only the bucket $g(q)$, but also the buckets $g(q) + \Delta$ that are likely to contain similar objects of q , are searched. Given the locality-sensitive of hash function (6), if an object similar to q is not in $g(q)$, then it is likely in the bucket with hash value slightly different from $g(q)$. Fig. 1 shows the distribution of bucket distances of N nearest neighbors. The left of Fig. 1 shows that for a query, more than 90% of the single hash values of its top- N nearest neighbors are either same as that of the query or differ by just -1 or 1, which indicates a majority of $\delta_k \in \{0, -1, 1\}$. The right of Fig. 1 shows that the probabilities of $\delta_k = 0$ are almost the same at each dimension. Hence, we restrict our attention to Δ with $\delta_k \in \{-1, 0, 1\}$. For a query q , its perturbation vectors can be derived based on the probability of finding a similar object of q in bucket $g(q) + \Delta$ (denoted by $\Pr(\Delta)$). Denote $\Pr_k(\delta_k)$ the probability of $(g(q) + \Delta)_k$ differs by δ_k from $g(q)_k$ and assume all the hash dimensions are independent, we have:

$$\Pr(\Delta) = \prod_{k=1}^K \Pr_k(\delta_k), \delta_k \in \{-1, 0, 1\} . \tag{7}$$

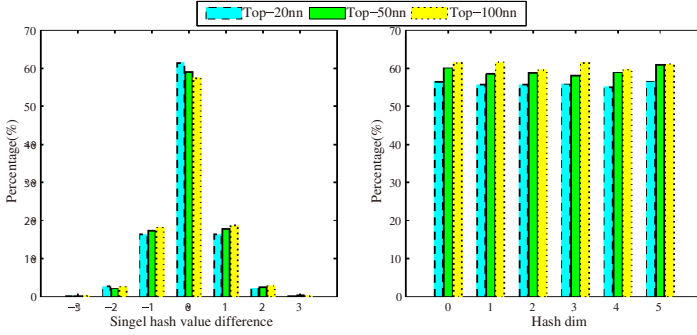


Fig. 1. Distribution of the bucket distances. The dataset used for illustration is SIFT1M [23], hash dim $K = 6$. The left is the distribution of single hash value differences (δ_k), the right is the probability of each hash dimension remains unchanged in perturbation vectors.

Normalize each $\Pr(\Delta)$ by dividing $\Pr(\mathbf{0}) = \prod_k \Pr_k(0)$ and take the negative logarithm of $\Pr(\Delta) / \Pr(\mathbf{0})$, (7) can be rewritten as:

$$\log_score(\Delta) = \sum_{k=1, \delta_k \neq 0}^K -\log \Pr'_k(\delta_k) . \tag{8}$$

where $\Pr'_k(\delta_k) = \Pr_k(\delta_k) / \Pr_k(0)$, $\Pr'_k(-1) + \Pr'_k(1) = 1 / \Pr_k(0) - 1$.

$\log_score(\Delta)$ only depends on the non-zero dimensions of Δ . Therefore, without taking account of $\Pr_k(0)$ as [8] dose, we define a score function to measure the probability of finding nearest neighbor of q in $g(q) + \Delta$:

$$score(\Delta) = \sum_{k=1, \delta_k \neq 0}^K -\log(p_k(\delta_k)) . \tag{9}$$

where $p_k(\delta_k) \propto \Pr'_k(\delta_k) \propto \Pr_k(\delta_k)$. In $score(\Delta)$, only $p_k(\delta_k)$ with $\delta_k \neq 0$ are summed, which means in $g(q) + \Delta$, only the dimensions different from $g(q)$ make contribution to $score(\Delta)$. Hence, $p_k(\delta_k)$ can be considered as a “generative probability”, and in this case we can set $p_k(-1) + p_k(1) = 1$. In the following, we show how to get a proper estimation of $p_k(\delta_k)$ with a simple method.

Fig. 2 illustrates the probability of q 's nearest neighbors falling into the neighboring slots. Here $f_k(q) = \omega_k^T q$, $h_k(q)$ is the slot to which q maps. W_{ku} , W_{kl} are the upper and lower boundaries of slot $h_k(q)$ respectively. Let $x_k(-1) = CDF_k(f_k(q)) - CDF_k(W_{kl})$, $x_k(1) = CDF_k(W_{ku}) - CDF_k(f_k(q))$. Obviously, the larger $x_k(-1)$ is, the larger $\Pr_k(1)$ is; the larger $x_k(1)$ is, the larger $\Pr_k(-1)$ is. Since $\Pr_k(-1) + \Pr_k(1)$ and $x_k(-1) + x_k(1)$ (which is $1/L$, L is the number of slots) are both constants, a reasonable assumption of the linear relationship between $x_k(\delta_k)$ and $\Pr_k(\delta_k)$ is: $\Pr_k(1) \propto x_k(-1)$; $\Pr_k(-1) \propto x_k(1)$. Moreover, we also have $p_k(-1) + p_k(1) = 1$ and $p_k(\delta_k) \propto \Pr_k(\delta_k)$. Therefore, based on the above derivations and approximations, we get a reasonable estimation of $p_k(\delta_k)$:

$$p_k(-1) = \frac{x_k(1)}{x_k(1) + x_k(-1)}, \quad p_k(1) = \frac{x_k(-1)}{x_k(1) + x_k(-1)}. \quad (10)$$

Given a query q , its corresponding $p_k(-1)$ and $p_k(1)$ for each dimension k is:

$$p_k(-1) = h_k(q) + 1 - L \cdot \text{CDF}_k(\omega_k^T q), \quad p_k(1) = 1 - p_k(-1). \quad (11)$$

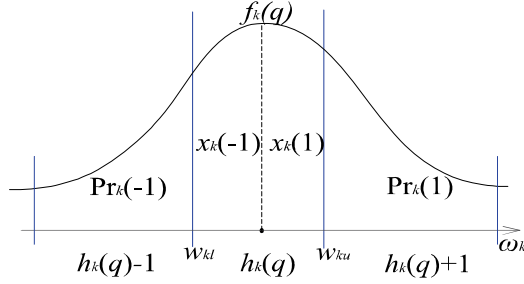


Fig. 2. Illustration of the probability of q 's nearest neighbors falling into the neighboring slots

The mathematical form of $\text{score}(\Delta)$ (9) is exactly the same as the score function defined in [8], hence, we could use the algorithms proposed in [8] to derive the perturbation vectors. After deriving perturbation vectors Δ , buckets $g(q) + \Delta$ are searched according to the ascending order of $\text{score}(\Delta)$. The complexity of deriving T perturbation vectors is almost $O(K \log K + \sum_{i=1}^T 2 \log i) < O(K \log K + T \log T)$. Note that, some approximations are made to simplify the derivation of Δ , and this may cause the final perturbation vectors non-optimal. However, the experimental results in Section 3 demonstrate the effectiveness of our multi-probe strategy, which indicate the efficacy of the derived perturbation vectors.

3 Experimental Result

Two public high-dimensional datasets [23] are used in our experiments, one consists of 1M 128-dimensional local SIFT features [24] (SIFT1M) and the other one consists of 1M 960-dimensional global GIST features [25] (GIST1M). Each dataset contains three vector subsets (database, query and learning) and a groundtruth set. For SIFT1M, the data points of each vector subset are normalized by dividing each dimension by the largest l_2 norm in the database set.

3.1 Performance Metric

Search accuracy and *Search complexity* are used to measure the similarity search performance of our method. These two measures reflect the objective which is of interest in many practical applications.

Search Accuracy. For ANN query, *Recall* (12) is used as the performance metric. It is the percentage of queries that the true nearest neighbor is in the result list [16].

$$\text{Recall} = \frac{1}{|Q|} \sum_{q \in Q} \delta(q) . \quad (12)$$

$\delta(q)$ is 1 if q 's nearest neighbor is in the result list or 0 if not. For k-Nearest Neighbor Search (kNN), distance *error ratio* [8] is used as the performance metric:

$$\text{error ratio} = \frac{1}{K|Q|} \sum_{q \in Q} \sum_{k=1}^K \frac{d_k - d_k^*}{d_k^*} . \quad (13)$$

where d_k is the distance between q and its k -th nearest neighbor in the result list, d_k^* is the distance between q and its true k -th nearest neighbor.

Search Complexity. The complexity of search process is of concern to many practical applications. Since the search process of our algorithm is similar to that of [16], the search complexity model in [16] is applicable to our algorithm. Two major phrases for search process are defined in [16] and listed below:

Phrase 1: query preparation cost (*qpc*). It measures the complexity of identifying the buckets that the search process will subsequently analyze in detail.

Phrase 2: short-list processing cost: *selectivity* (*sel*). It is the fraction of the number of data points in the result list to the number of data points in total dataset.

The overall search cost of our algorithm is: $\text{sel} \cdot nd + \text{qpc}$. The acceleration factor *ac* [16] defined in (14) is used to measure the speedup over linear search:

$$\text{ac} = \frac{nd}{\text{sel} \cdot nd + \text{qpc}} = \frac{1}{\text{sel} + \text{qpc}/(nd)} . \quad (14)$$

3.2 Experimental Setup

Label Matrix Construction. We randomly sample l points from the learning set to form the neighbor and non-neighbor pairs. The distance between each data point p in the sample set \mathbf{X}_l and its 150th nearest neighbor is calculated and averaged (\bar{d}_{150}). For any $x_i, x_j \in \mathbf{X}_l$: if $d(x_i, x_j) \leq \bar{d}_{150}$, (x_i, x_j) is a neighbor pair; if $d(x_i, x_j) \geq t\bar{d}_{150}$, (x_i, x_j) is a non-neighbor pair; otherwise, the relationship of x_i and x_j is not defined. For SIFT1M, $l=2000$, $t=1.5$, and for GIST1M, $l=5000$, $t=2.5$.

Parameters Settings. There are three key parameters influencing the performance of DALSH: the dimensionality of hash value K , the number of slots of each hash dimension L (L_k is set to the same value) and the number of Gaussian components g_k . K and L are varied in our experiments to get different performance. The KL divergence between the actual and estimated distributions is used to adjust g_k . With a larger g_k , the estimation is more accurate. However, the performance is not improved

much when g_k exceeds 4, moreover, a larger g_k may also lead to overfitting. Therefore, in our experiments, g_k is set to 3 in the experiments on dataset SIFT1M and 4 on dataset GIST1M. The experimental results are averaged over all queries (10,000 for SIFT1M and 1,000 for GIST1M) with one single hash table.

3.3 Comparisons with “original” LSH Methods

In [16], Paulevé et al. have evaluated several hash functions of LSH by comparing Recall (12) of each hash function at a given selectivity. They argue that the k-means based LSH (KLSH) gives the best result. Since recall at a given selectivity reflects the effectiveness of a hash function, we do the same evaluation for DALSH with one single hash table. Fig. 3 gives the evaluation of distribution-aware hashing (6) compared with several other hashing schemes evaluated in [16]¹. Note that, each point in Fig. 3 corresponds to an optimal parameters setting that gives the best performance for a given selectivity. As can be seen from this figure, the distribution-aware hashing significantly outperforms the random projection [8] and lattice [15] based hashing: the selectivity is almost one order of magnitude smaller for the same recall. Moreover, the performance of distribution-aware hashing is almost as good as (even better than) that of the k-means based hashing (k-means, HKM) when recall is larger than 40%.

We evaluate the performance of DALSH for similarity search compared with the popularly used E2LSH and the state-of-the-art “original” LSH method, Kmeans LSH (KLSH). We report the acceleration factor (14) of each method at the same Recall. In DALSH, the complexity of Phrase 1 is almost $O(K \log K + T \log T) + O(dK)$. Since $n, d \gg K, nd \gg T \log T$, the acceleration is almost $1/\text{sel}$. The top row of Fig. 4 gives the experimental results for ANN query on SIFT1M and GIST1M respectively. As shown, our method gives the best performance on both two datasets: its recall is almost 10%-20% higher than those of E2LSH and KLSH, and its query speed is almost 6 and 1.2-4 times faster than those of E2LSH and KLSH respectively.

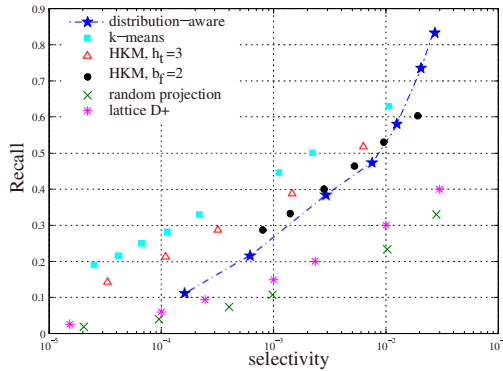


Fig. 3. Evaluation of the distribution-aware hash function on SIFT1M, compared with the (hierarchical) k-means, random projection and lattice based hash functions evaluated in [16]

¹ The dataset used in our experiment is the same as [16], thus the experimental results of us and [16] can be compared directly. Only some representative points of the “selectivity-recall” curve in [16] are depicted, for more details, please refer to Fig. 2 of [16].

For kNN query, the error ratio (13) is used as the performance metric and we set $K=100$. As shown in the bottom row of Fig. 4, DALSH and KLSH give almost the same performance (both outperform E2LSH), their query speed are both almost 10 times faster than that of E2LSH. It is not surprising that KLSH gives the best performance for kNN query, since the inner distance of each cluster is minimized when clustering. However, in this case, the acceleration of KLSH is relatively small.

3.4 More Comparisons

We also compare DALSH with several other kinds of high-dimensional index methods, such as optimized KD-tree [26], FALNN [27] and spectral hashing [18]. The experimental results are succinctly presented below due to the space limitation.

When compared with spectral hashing (SpH), the code length is set to 16, 32 and 64, and the query results are those with hamming distance to the query less than 6. SpH's performance is not improved much with a longer code and the highest recall is almost 40% on SIFT1M, which is inferior to that of DALSH.

The optimized KD-tree is implemented within FLANN. We change the parameter “algorithm” and “target_precision” to get different performance. In the experiments on SIFT1M, the best recall of optimized KD-tree is 67.35% with a speedup less than 100, both are inferior to those of DALSH. The best performance of FALNN is obtained by setting “algorithm” to “FLANN_INDEX_AUTOTUNED”, and in this case, the recall is 80.45% and the speedup is 77.2, while the speedup of DALSH is almost 200 at the same recall. On GIST1M, since the dimension is high (960), FLANN's performance degrades greatly, which is consistent with the result in [27], while DALSH's performance is almost unchanged (see. Fig. 4).

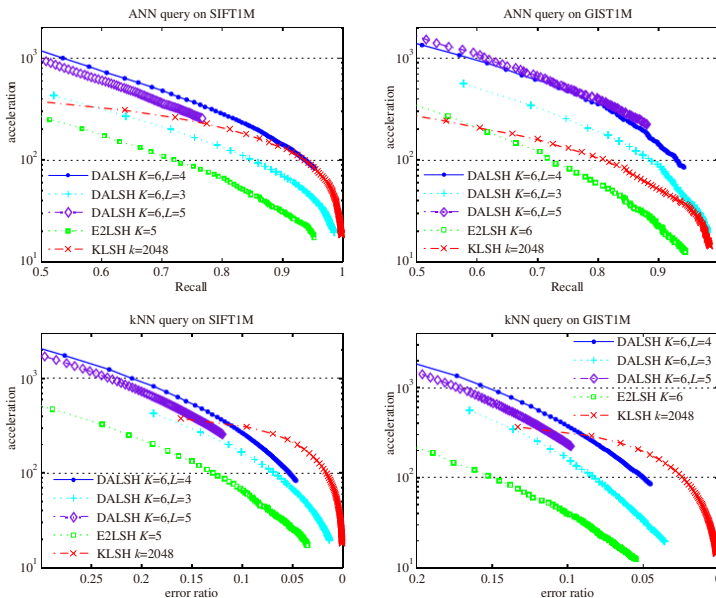


Fig. 4. Evaluations of DALSH, compared with E2LSH and KLSH. L is the number of slots in DALSH, K is the hash dimension, k is the number of clusters in KLSH. The top and bottom row give the acceleration of DALSH over linear search for ANN and kNN query respectively.

4 Conclusions

In this paper, we propose the Distribution-Aware LSH (DALSH) for high-dimensional similarity search. With a supervised learning algorithm, we generate a series of data-adaptive projection vectors, each of which tries to minimize the difference between the projections of similar objects, while maximize the difference between the projections of dissimilar objects on it. Linear projection is used to reduce the dimensionality of the dataset to index, and then the hash functions of DALSH are derived from the distribution of the dimension-reduced data. In this way, the problem of lacking adaptation to real data, which is the intrinsic limitation of most existing “original” LSH methods, is alleviated. In addition, we present an efficient multi-probe strategy to improve the query performance of DALSH. The experimental results on two public high-dimensional datasets demonstrate the efficacy of DALSH. Compared with the state-of-the-art “original” LSH method, Kmeans LSH (KLSH), DALSH shows commendable performance gains in terms of search accuracy and search efficiency. Moreover, compared with other kinds of index methods (e.g. optimized KD-tree, spectral hashing), DALSH still gives a better performance.

Acknowledgments. This work is supported by National Nature Science Foundation of China (61273247, 61271428), National Key Technology Research and Development Program of China (2012BAH39B02), and Co-building Program of Beijing Municipal Education.

References

1. Bentley, J.L.: K-d trees for semidynamic point sets. In: Proc. SCG, pp. 187–197 (1990)
2. Guttman, A.: R-trees: a dynamic index structure for spatial searching. In: International Conference on Management of Data, pp. 47–57 (1984)
3. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is “nearest neighbor” meaningful? In: International Conference on Database Theory, pp. 217–235 (1999)
4. Tao, Y., Yi, K., Sheng, C., Kalnis, P.: Efficient and accurate nearest neighbor and closest pair search in high-dimensional space. ACM TODS 35(3), 20 (2010)
5. Gionis, A., Indyk, P., Motwani, R.: Similarity search in high dimensions via hashing. In: Proceedings of the International Conference on Very Large Data Bases, pp. 518–529 (1999)
6. Indyk, P., Motwani, R.: Approximate nearest neighbors: towards removing the curse of dimensionality. In: Proc. STOC, pp. 604–613 (1998)
7. Datar, M., Immorlica, N., Indyk, P., Mirrokni, V.S.: Locality-sensitive hashing scheme based on p-stable distributions. In: Proc. SCG, pp. 253–262 (2004)
8. Lv, Q., Josephson, W., Wang, Z., Charikar, M., Li, K.: Multi-probe lsh: efficient indexing for high-dimensional similarity search. In: Proc. VLDB, pp. 950–961 (2007)
9. Wang, J., Kumar, S., Chang, S.F.: Sequential projection learning for hashing with compact codes. In: Proc. ICML, pp. 1127–1134 (2010)
10. Kulis, B., Grauman, K.: Kernelized locality-sensitive hashing for scale image search. In: International Conference on Computer Vision, pp. 2130–2137 (2009)

11. Liu, W., Wang, J., Kumar, S., Chang, S.F.: Hashing with graphs. In: ICML, pp. 1–8 (2011)
12. Joly, A., Buisson, O.: A posteriori multi-probe locality sensitive hashing. In: ACM MM (2008)
13. Bawa, M., Condie, T., Ganesan, P.: LSH forest: self-tuning indexes for similarity search. In: Proc. WWW, pp. 651–660 (2005)
14. Shakhnarovich, G., Darrell, T., Indyk, P.: Nearest-neighbor methods in learning and vision. *IEEE Transactions on Neural Networks* 19(2), 337 (2008)
15. Jégou, H., Amsaleg, L., Schmid, C., Gros, P.: Query adaptative locality sensitive hashing. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 825–828 (2008)
16. Paulevé, L., Jégou, H., Amsaleg, L.: Locality sensitive hashing: A comparison of hash function types and querying mechanisms. *Pattern Recognition Letters* 31(11), 1348 (2010)
17. Salakhutdinov, R., Hinton, G.: Semantic hashing. *Int. J. Approx Reason.* 50(7), 969 (2009)
18. Weiss, Y., Torralba, A., Fergus, R.: Spectral hashing. In: NIPS, pp. 1753–1760 (2008)
19. Raginsky, M., Lazebnik, S.: Locality-sensitive binary codes from shift-invariant kernels. In: Neural Information Processing Systems, pp. 1509–1517 (2009)
20. Wang, M., Yang, K., Hua, X., Zhang, H.: Towards a relevant and diverse search of social images. *IEEE Transactions on Multimedia*, 829–842 (2010)
21. Liu, W., Wang, J., Ji, R., Jiang, Y.G., Chang, S.F.: Supervised hashing with kernels. In: Proc. CVPR, pp. 2074–2081 (2012)
22. Lejsek, H., Ásmundsson, F.H., Jónsson, B.T., Amsaleg, L.: NV-tree: An efficient disk based index for approximate search in very large high-dimensional collections. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(5), 869 (2009)
23. Jégou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(1), 117 (2011)
24. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 91–110 (2004)
25. 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 (2001)
26. Silpa-Anan, C., Hartley, R.: Optimised KD-trees for fast image descriptor matching. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
27. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: Int. Conf. Computer Vision Theory and Applications, pp. 331–340 (2009)

Cross Concept Local Fisher Discriminant Analysis for Image Classification

Xinhang Song^{1,2}, Shuqiang Jiang^{1,2}, Shuhui Wang^{1,2}, Jinhui Tang³,
and Qingming Huang^{1,2,4}

¹ Key Lab of Intell. Info. Process, Chinese Academy of Sciences, Beijing 100190, China

² Inst. of Comp. Tech., Chinese Academy of Sciences, Beijing 100190, China

³ School of Comp. Sci. and Tech., Nanjing Univ. of Sci. and Tech.,
Nanjing 210093, China

⁴ University of Chinese Academy of Sciences, Beijing 100049, China 100049, China
{xhsong,sqjiang,shwang,qmhuang}@jdl.ac.cn, jinhuitang@mail.njust.edu.cn

Abstract. Distance metric learning is widely used in many visual computing methods, especially image classification. Among various metric learning approaches, Fisher Discriminant Analysis (FDA) is a classical metric learning approach utilizing the pair-wise semantic similarity and dissimilarity in image classification. Moreover, Local Fisher Discriminant Analysis (LFDA) takes advantage of local data structure in FDA and achieves better performance. Both FDA and LFDA can only deal with images with simple concept relations, where images either belong to the same concept category or come from different categories. However, in real application scenarios, images usually contain multiple concepts, and relations of concepts and images are complex. In this paper, to improve the flexibility of LFDA on the complex image-concept relations, we propose a new pairwise constraints method called Cross Concept Local Fisher Discriminant Analysis (C^2 LFDA) for image classification. By considering the cross concept images as a special case of within-class samples, C^2 LFDA models the semantic relations of images for distance metric learning under the framework of LFDA. We calculate within-class and between-class scatter matrix based on the proposed re-weighting scheme and local manifold structure. By solving the objective function of discriminant analysis using the proposed scheme, a set of projected representation is obtained to better reflect the complex semantic relations among images. Experimental evaluations and comparisons show the effectiveness of the proposed method.

Keywords: Distance metric learning, Multiple concepts, Fisher Discriminant Analysis.

1 Introduction

Distance metric is a crucial issue in visual computing, which serves an important role in image retrieval and image classification. Distances can be directly used for

unsupervised clustering - such as spectral methods for example, or for supervised classification - such as nearest neighbor classification [1]. Compared with the direct distance computing methods, the main goal of distance metric learning (DML) is to make the processed data having a better ability on compactness and semantic consistency. DML has been intensively investigated in the literature and it is a useful way to differentiate images learning problems with different semantic information.

Supervised distance metric learning is to learn a distance metric according to images' label information. Among the existing approaches [2-4] are traditional supervised distance metric learning methods by using pairwise constraints. Specially, their works [5-8] are distance metric learning methods that can maintain samples' local neighborhood structure according to the visual distance between samples.

The above mentioned methods are only capable of data with simple concept relations. However, the situation is not always like that, there might be several concepts in one image. Two images may belong to one category for the same label, yet they may have other concepts that do not belong to one category [9, 10]. Researchers have paid attentions to this problem and several solutions have been proposed, which can be grouped into two main categories: a) problem transformation methods [11], and b) algorithm adaptation methods [12, 13]. These two methods either transform multi-label data into single-label or find the optimal values through methods based on SVM or Boost. These methods can not reflect the label correlations, their works [14, 15] make use of the label correlations to improve classification accuracy. Specially in [14], they propose a multi-label multi-class classification method based on LDA (Linear Discriminant Analysis). They compute label correlation statistics, and then use this label correlation in the training procedure of multi-label LDA. Although their method make use of the label correlation, they didn't consider the local structure between samples.

In this paper, we propose a method called Cross Concept Local Fisher Discriminant Analysis (C^2 LFDA) which can describe the similarities in both visual and semantic domain of training data. The visual similarity is the similarity in visual feature space. And the semantic similarity means the label correlation of multiple concepts data. Our method deal with image data as shown in Fig. 1 which not only contain simple concept relations but also multiple and complex concepts and we also call them label overlapping data. We assign label overlapping data to each associated class according to the corresponding labels. Then we redefine the within-class and between-class scatter matrix. Thus label overlapping data will be calculated in each associated class. However the label overlapping data is different with simple concept data, we can not train them by directly using LFDA. We re-weight the within-class and between-class scatter matrix by a weight which can reflect the influential factor of label overlapping data in each associated class. Then we calculate the weighted within-class and between-class scatter matrix based on the proposed re-weighting scheme and local manifold structure. And the target transformation matrix can be obtained by

solving an object function based on our re-weighted within-class and between-class scatter matrix.

Our contributions in this paper are: first, we propose a distance metric learning method called C²LFDA which can deal with images with multiple and complex concepts; second, by re-weighting the within-class and between-class scatter matrix with the similarities in both visual and semantic domain, we can get a better classification performance than LFDA.

The rest of paper is organized as follows. In section 2, we discuss related works. In section 3, we briefly review FDA and LFDA. In section 4, we define C²LFDA and show its fundamental properties. In section 5, we compare C²LFDA with LFDA and Euclidean distance for the task of image classification, and we obtain promising results. Finally, we give concluding remarks and future prospects in Section 6.

2 Preliminaries

2.1 Formulation

Let $x_i \in \mathbb{R}^d$ ($i = 1, 2, \dots, n$) be d -dimensional samples and X be the matrix of all samples:

$$X \equiv (x_1|x_2|\dots|x_n) \tag{1}$$

Where n is the number of samples.

Let $z_i \in \mathbb{R}^r$ ($1 \leq n \leq d$) be embedded samples, where n is the dimension of embedding space. We focus on classification for this moment, i.e., using a $d \times r$ transformation matrix T . So the embedded space z can be represented as:

$$z_i = T^\top x_i \tag{2}$$

2.2 Fisher Discriminant Analysis (FDA)

Here we briefly review the definition of Fisher discriminant analysis(FDA) [16] [2] [5].

Let c be the number of labels and $y_i \in \{1, 2, \dots, c\}$ be a class label associated with the sample x_i . Let n_j be the number of labeled samples in class j .

Let $S^{(w)}$ and $S^{(b)}$ be the within-class scatter matrix and the between-class scatter matrix:

$$S^{(w)} = \sum_j^c \sum_{i:y_i=j} (x_i - u_j)(x_i - u_j)^\top \tag{3}$$

$$S^{(b)} = \sum_{i:y_i=j} n_i(x_i - u_j)(x_i - u_j)^\top \tag{4}$$

where $u_j \equiv \frac{1}{n_j} \sum_{i:y_i=j} x_i$ and $\mu \equiv \frac{1}{n} \sum_{i=1}^n x_i$ Using $S^{(w)}$ and $S^{(b)}$ the FDA transformation matrix T_{FDA} is defined as follows:

$$T_{FDA} = \operatorname{argmax}_{T \in \mathbb{R}^{d \times c}} \operatorname{tr}((T^\top S^{(w)} T)^{-1} T^\top S^{(b)} T) \tag{5}$$

That is, we can seek a transformation matrix T such that the between-class scatter is maximized while the within-class scatter is minimized. Then a solution of T_{FDA} is given by

$$T_{FDA} = (\varphi_1|\varphi_2|\dots|\varphi_c) \quad (6)$$

Where $\{\varphi_i\}_{i=1}^d$ are the generalized eigenvectors associated to the generalized eigenvalues $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_d$ of the following generalized eigenvalue problem:

$$S^{(b)}\varphi = \lambda S^{(w)}\varphi \quad (7)$$

The between-class scatter matrix $S^{(b)}$ has at most rank $c-1$ [2], thus FDA can find at most $c-1$ meaningful features which is the limitation of FDA.

2.3 Local Fisher Discriminant Analysis (LFDA)

Local Fisher discriminant analysis (LFDA) overcomes vulnerability of original FDA against within-class multimodality or outliers [17].

Let $S^{(lw)}$ and $S^{(lb)}$ be the local between-class scatter matrix and the local within-class scatter matrix defined by:

$$S^{(lw)} = \frac{1}{2} \sum_{i,j=1}^n W_{i,j}^{(w)} (x_i - x_j)(x_i - x_j)^\top \quad (8)$$

$$S^{(lb)} = \frac{1}{2} \sum_{i,j=1}^n W_{i,j}^{(b)} (x_i - x_j)(x_i - x_j)^\top \quad (9)$$

Where:

$$W_{i,j}^{(w)} = \begin{cases} A_{i,j}/n_k & \text{if } y_i = y_j = k \\ 0 & \text{if } y_i \neq y_j \end{cases} \quad (10)$$

$$W_{i,j}^{(b)} = \begin{cases} A_{i,j}(1/n - 1/n_k) & \text{if } y_i = y_j = k \\ 1/n & \text{if } y_i \neq y_j \end{cases} \quad (11)$$

This weight the values for the sample pairs in the same class according to the affinity matrix A . Thus, LFDA seeks a transformation matrix T which has the following properties: 1) nearby data pairs in the same class are made close and the data pairs in different classes are made apart; 2) far apart data pairs in the same class are not imposed to be close. Samples in different classes are separated from each other irrespective of their affinity values. A solution T_{LFDA} is can be obtained like FDA

$$S^{(lb)}\varphi = \lambda S^{(lw)}\varphi \quad (12)$$

The local between-class $S^{(lb)}$ generally has a much higher rank with less eigenvalue multiplicity because of the local factor $A_{i,j}$ [6].

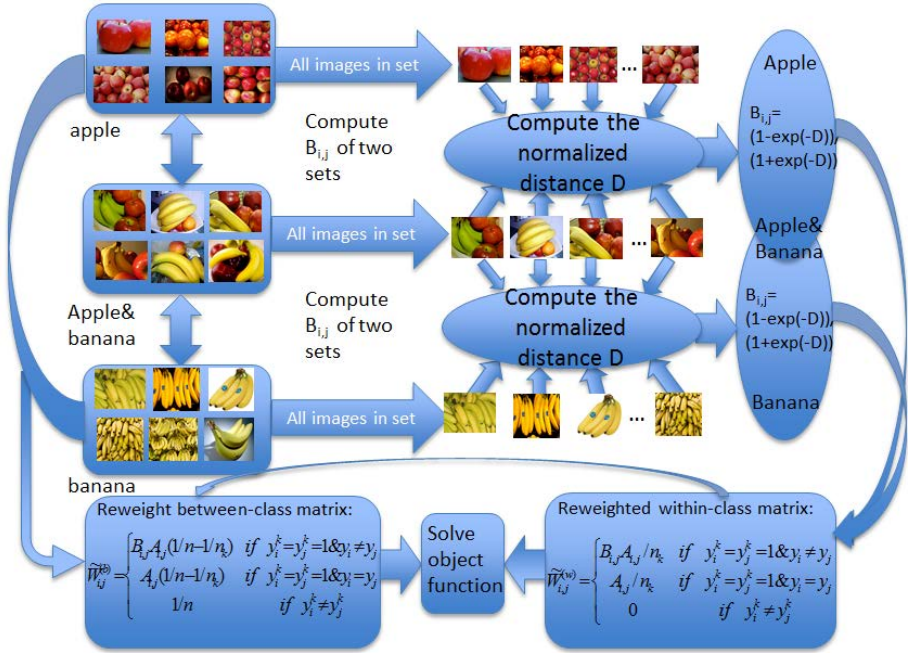


Fig. 1. These are three sets of images from our dataset. The set contains apple and banana concept is within-class with images in apple set as well as the images in banana set. Yet the relationship of within-class cannot be transmitted, which is to say, apple set and banana set are not within-class because of this, they are between-class. We re-weight the within-class and the between-class scatter matrix by using the similarity (affinity matrix A and B) in both visual and semantic domain. We can get the matrix element $B_{i,j}$ between two sets i and set j through the following procedure: first, calculate the distances that form all samples in set i to all samples in set j ; then normalize all these distances and get the mean of this distance: d ; finally $B_{i,j} = (1 - \exp(-d)) / (1 + \exp(-d))$. Affinity matrix A gets from LPP. We get the target transformation matrix by solving an object function $T_{C^2LFDA} = \operatorname{argmax}_{TR^{d \times c}} \operatorname{tr}((T^T S^{(w)} T)^{-1} T^T S^{(b)} T)$.

3 Cross Concept Fisher Discriminant Analysis

3.1 Basic Idea

Conventional supervised distance learning methods learn a distance metric according to images' labels, and they require every sample has only one label, which are contradicted by the actual facts. As a matter of fact, samples usually have more than one label, which may affect the training results inevitably if we only use one label in the sample. Therefore, what we want to do is to use all the label information of the image comprehensively. We divide images containing same label into one category, so there must be an intersection between two categories. For example, there is a set of images U containing label A and label

B; $I(A)$ and $I(B)$ represent all the images category A and B, so the intersection of category A and category B is set $U=I(A)\cap I(B)$. The images included in the intersection U may not help in two-notion classification between category A and B, but in multi-class classification between A,B and others. Thus, our method in this paper is to make use of these image sets like U on the framework of LFDA training method. Images intersection set U contains multiple concepts, the importance of every label will not be the same in different feature space. In order to use images U precisely, we apply different coefficients to distinguish the importance of different concepts in training.

To be more specific, LFDA minimize within-class distances and maximize between-class distances in training, however, in this method, there is a little bit difference when calculating between-class distance and within-class distance comparing with the former ones. This is because image intersection U has multiple labels. These images in U are involved in the calculation of all the associated within-class; and they are exempted from the calculation of between-class distances. More specific formularized expressions will be given in the next section.

3.2 Definition

In LFDA, the formula (8) (9) use affinity matrix $A_{i,j}$ to weight pairs within classes. Meanwhile, we need to weight different classes, so we use affinity matrix B. Finally, the original $W_{i,j}^{(w)}$ and $W_{i,j}^{(b)}$ turn out to be:

$$\tilde{W}_{i,j}^{(b)} = \begin{cases} B_{i,j}A_{i,j}/n_k & \text{if } y_i^k = y_j^k = 1 \& y_i \neq y_j \\ A_{i,j}/n_k & \text{if } y_i^k = y_j^k = 1 \& y_i = y_j \\ 0 & \text{if } y_i^k \neq y_j^k \end{cases} \quad (13)$$

$$\tilde{W}_{i,j}^{(w)} = \begin{cases} B_{i,j}A_{i,j}(1/n - 1/n_k) & \text{if } y_i^k = y_j^k = 1 \& y_i \neq y_j \\ A_{i,j}(1/n - 1/n_k) & \text{if } y_i^k = y_j^k = 1 \& y_i = y_j \\ 1/n & \text{if } y_i^k \neq y_j^k \end{cases} \quad (14)$$

Where $y_i \in \{0, 1\}^c$ is a binary vector. $y_i^k=1$ means that sample x_i has the k_{th} label, otherwise not. And here affinity matrix $B_{i,j}$ represents the similarity between the class corresponding to sample i and j . Now, within-class and between-class $\tilde{W}_{i,j}^{(w)}$, $\tilde{W}_{i,j}^{(b)}$, turn out to be:

$$\tilde{S}^{(w)} = \frac{1}{2} \sum_{i,j=1}^n \tilde{W}_{i,j}^{(w)} (x_i - x_j)(x_i - x_j)^\top \quad (15)$$

$$\tilde{S}^{(b)} = \frac{1}{2} \sum_{i,j=1}^n \tilde{W}_{i,j}^{(b)} (x_i - x_j)(x_i - x_j)^\top \quad (16)$$

Using $\tilde{S}^{(w)}$ and $\tilde{S}^{(b)}$ the C^2LFDA transformation matrix T_{C^2LFDA} is defined as follows:

$$T_{C^2LFDA} = \operatorname{argmax}_{T \in R^{d \times c}} \operatorname{tr}((T^\top \tilde{S}^{(w)} T)^{-1} T^\top \tilde{S}^{(b)} T) \quad (17)$$

That is, we can seek a transformation matrix T such that the between-class scatter is maximized while the within-class scatter is minimized.

Then a solution of T_{C^2LFDA} is given by

$$T_{C^2LFDA} = (\varphi_1|\varphi_2|\dots|\varphi_c) \quad (18)$$

Where $\{\varphi_i\}_{i=1}^d$ are the generalized eigenvectors associated to the generalized eigenvalues $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_d$ of the following generalized eigenvalue problem:

$$\tilde{S}^{(b)}\varphi = \lambda\tilde{S}^{(w)}\varphi \quad (19)$$

3.3 Properties

Our method is to apply the traditional LFDA to label overlapping distance metric learning with the similarity between multiple image sets and the other sets. It needs to redefine within-class and between-class scatter matrix when applying LFDA to multiple label data, which is to say if one pair of samples have the same label, we consider the relationship between this pair as within class of this label, otherwise between-class. In this case, within-class pairs of samples must have some same label, though they may have different labels. In other words, the relationship between sample labels is not like what it is in the original LFDA, which is either the same or different.

Since relationship between labels of the samples becomes more complicated, we can use similarity between multi-label sets and other sets to describe the relationship further. This similarity can represent multi-label sets' effect as well. We use affinity matrix B to represent this similarity. The matrix element $B_{i,j}$ will be the similarity between set i and set j and it can measure influential factor between them. Here set i is multi-label set and set j is the corresponding set with i 's label. A higher $B_{i,j}$ represents that set i and set j affect each other more than other sets in within-class scatter class. It will not work if we set all $B_{i,j}$ with value 1. So a proper method to compute affinity matrix B will work better. It is effective in C^2LFDA when combining these two similarities which has been verified in the experiment results in next section.

4 Experiments

4.1 Dataset

As to our method, we collect labeled data including 45 sets 30 classes and 11294 samples. Three of them are selected from dataset shown in Fig. 1. We try to utilize the relations between label overlapping image sets and other simple concept image sets, therefore we put those multiple concepts images which share same labels into same class when we collect data.

4.2 Distance Metric Learning for Classification

The main idea of testing our method is to learn a transform matrix T with training samples, according to which we calculate the Mahalanobis distance for testing samples. Then we can use KNN classifier to classify the testing samples, getting the accuracy rate of this method and original LFDA.

In our method, the affinity matrix B represents the similarity between the set of label overlapping images and their associated class according to their labels. We can get the matrix element $B_{i,j}$ in Eq.[13, 14] between two sets i and j through the following procedures: first, calculate the distances that form all samples in set i to all samples in set j ; then normalize all these distances and get the mean of this distance: d ; finally $B_{i,j} = (1 - \exp(-d)) / (1 + \exp(-d))$. This indicates the similarity between sets, similar and sharing same labels sets become much closer.

4.3 Comparison

We use three kind of features on three ways to measure the distance: 1)KNN: we calculate the Euclidean distance directly then we use KNN. As shown in Fig. 2 (a); 2)LFDA: we use LFDA to learn a distance metric and then apply KNN. As shown in Fig. 2 (b); 3)C²LFDA: we use C²LFDA to measure and same as above. It is shown in Fig. 2 (c).

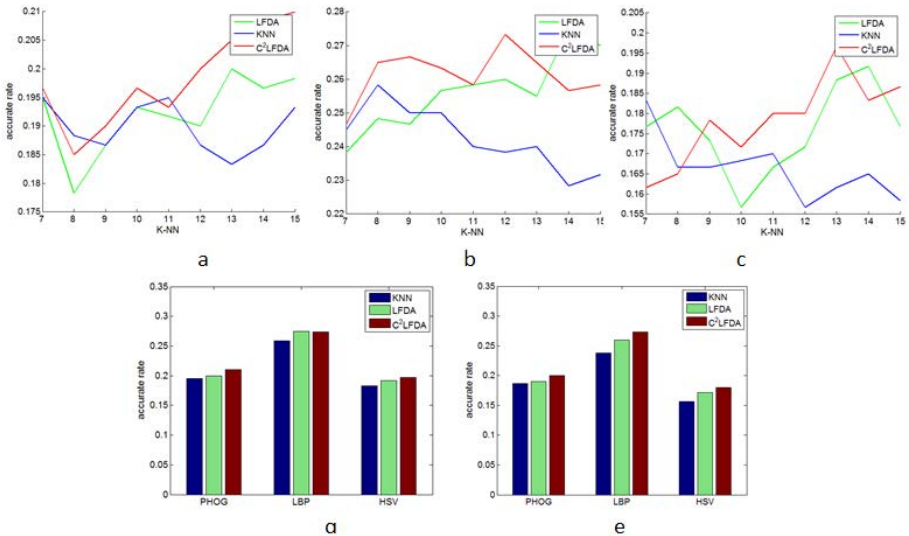


Fig. 2. Experiment results: (a), (b) and (c) are the experiment result by PHOG, LBP and HSV respectively. The horizontal ordinate x represents the K from KNN. (d) is the optimal of the above three images. (e) is the result when $K=12$.

It can be observed from Fig. 2 (a)(b)(c) that the result of LFDA is better than original Euclidean distance in most cases, yet this method is generally superior to LFDA. Fig. 2 (d) is the statistic of optimal results of these three different methods using three features. It is clear that our optimal result is better than both LFDA and Euclidean distance. However, in practical applications, the K in KNN is usually a fixed value. It is shown in Fig. 2 (e), we can get the best result in each feature when $K=12$.

5 Conclusions

This work is focusing on how to use the relationship between different sets of images for distance metric learning. In this paper, we proposed a new method called C^2 LFDA, which redefines within-class and between-class matrix in LFDA and assign label overlapping samples into their associated classes. To be more specific, in within-class scatter matrix from C^2 LFDA samples are within-class if only some label of the samples is the same, otherwise they are between-class. Meanwhile, we put forward an affinity matrix B to represent the similarity between sets of images.

The distance matrix learned by LFDA trained with multiple concepts data is better than traditional Euclidean distance in KNN classifier. Yet our method C^2 LFDA with the affinity matrix B is better than LFDA. From the result of the experiment, promising results are achieved through our method. In the future work, we will investigate more effective methods to compute the affinity matrix B , in order to further improve the performance C^2 LFDA.

Acknowledgements. This work was supported in part by National Basic Research Program of China (973 Program):2012CB316400, in part by National Natural Science Foundation of China: 61070108,61025011, in part by The Natural Science Foundation of Jiangsu Province under Grant BK2012033 and BK2011700, and in part by Research Fund for the Doctoral Program of Higher Education of China (RFDP) under Grant 20113219120022.

References

1. Weinshall, D., Zamir, L.: Image Classification from Small Sample, with Distance Learning and Feature Selection. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Paragios, N., Tanveer, S.-M., Ju, T., Liu, Z., Coquillart, S., Cruz-Neira, C., Müller, T., Malzbender, T. (eds.) ISVC 2007, Part II. LNCS, vol. 4842, pp. 106–115. Springer, Heidelberg (2007)
2. Fukunaga, K.: Introduction to statistical pattern recognition (1990)
3. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. In: NIPS (2005)
4. Weinberger, K., Blitzer, J., Saul, L.: Distance metric learning for large margin nearest neighbor classification. In: NIPS, pp. 1475–1482 (2006)

5. He, X., Niyogi, P.: Locality preserving projections. In: Thrun, S., Saul, L., Scholkopf, B. (eds.) *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge (2004)
6. Sugiyama, M.: Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis. *Journal of Machine Learning Research* 176, 1027–1482 (2006)
7. Timofte, R., Van Gool, L.: Iterative nearest neighbors for classification and dimensionality reduction. In: *CVPR* (2012)
8. Hastie, T., Tibshirani, R.: Discriminant adaptive nearest neighbor classification. *Pattern Analysis and Machine Intelligence* 18
9. Tang, J., Zha, Z.J., Tao, D., Chua, T.S.: Semantic-gap oriented active learning for multi-label image annotation. *IEEE Transactions on Image Processing* 21
10. Li, L., Jiang, S., Huang, Q.: Learning hierarchical semantic description via mixed-norm regularization for image understanding. *IEEE Transactions on Multimedia* (2012)
11. Boutell, M., Luo, J., Shen, X., Brown, C.: Learning multi-label scene classification. *Pattern Recognition* 37, 1757–1771 (2004)
12. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. *Int. J. Data Warehousing and Mining 2007*, 1–13
13. Yang, L., Jin, R.: Distance metric learning: A literature survey. Michigan State University (2006)
14. Wang, H., Ding, C., Huang, H.: Multi-label Linear Discriminant Analysis. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010, Part VI*. LNCS, vol. 6316, pp. 126–139. Springer, Heidelberg (2010)
15. Wang, H., Huang, H., Ding, C.: Image annotation using bi-relational graph of images and semantic labels. In: *CVPR* (2011)
16. Fisher, R.: The use of multiple measurements in taxonomic problems. *Annals of Eugenics* (1936)
17. Sugiyama, M., Idé, T., Nakajima, S., Sese, J.: Semi-Supervised Local Fisher Discriminant Analysis for Dimensionality Reduction. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) *PAKDD 2008*. LNCS (LNAI), vol. 5012, pp. 333–344. Springer, Heidelberg (2008)

A Weighted One Class Collaborative Filtering with Content Topic Features

Ting Yuan¹, Jian Cheng¹, Xi Zhang¹, Qinshan Liu², and Hanqing Lu¹

¹ National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, Beijing 100190, China

² CICE, Nanjing University of Information Science and Technology, Nanjing 210044, China
{tyuan, jcheng, xi.zhang, luhq}@nlpr.ia.ac.cn, qslu@nuist.edu.cn

Abstract. A task that naturally emerges in recommender system is to improve user experience through personalized recommendations based on user's implicit feedback, such as news recommendation and scientific paper recommendation. Recommendations dealing with implicit feedback are most thought of as One Class Collaborative Filtering (OCCF), which only positive examples can be observed and the majority of data are missing. The idea to introduce weights for treating missing data as negatives has been shown to help in OCCF. But existing weighting approaches mainly use the statistical properties of feedback to determine the weight, which are not very reasonable and not personalized for each user-item pair. In this paper, we propose to improve recommendation by considering the rich user and item content information to assist weighting the unknown data in OCCF. To incorporate the useful content information, we get a content topic feature for each user and item by using probabilistic topic modeling method, and determine the personalized weight of every unknown user-item pair by these content topic features. Extensive experiments show that our algorithm can achieve better performance than the state-of-art methods.

Keywords: One-Class Collaborative Filtering, Recommender system, Implicit feedback, Topic modeling, Content topic feature.

1 Introduction

As living in the age of information explosion, one may find that it is increasingly difficult to explore the big data and sometimes even become confused about what he really wants. To overcome the information overload, recommender systems are attracting more and more attention. Nowadays recommender systems are not only widely used to recommend products in e-commerce websites such as Amazon and Netflix, what's more, they are also used to mine user's personalized preference of other useful large-scale information, such as news recommendation in Google news and scientific article recommendation in CiteUlike.

As one of the most popular methods in recommender system [1], Collaborative Filtering (CF) approaches aim at predicting the user's preference of items based on the rating history of all users. Neighborhood models [2, 3, 18] to make prediction by the performance of neighbors and latent factor models [8, 9] (also called matrix

factorization models) to decompose rating matrix by two low rank latent factors are the two basic approaches to CF, especially the latent factor models have achieved outstanding results in many recommendation competitions, such as the million-dollar Netflix competition¹. However, most CF methods focus on data sets with explicit ratings expressed in different scores (such as Netflix 1-5 scale scores). Such explicit ratings are not always available in practice. In real-world recommendation task, the data we deal with is often implicit feedback data, for example the users' click history in news recommendation, purchase history in products recommendation and post history in scientific articles recommendation. The biggest difference between explicit data and implicit data is that the implicit data has no negative feedback. For example, in Netflix data, if a user rates an item score 1, we can know the user does not like it, while in purchase history, if a user does not buy a product, we cannot determine whether he does not like it or even not see it. Thus, in those cases, all the observations belong to a single positive class, which leads to a one-class problem.

Recently, One-Class Collaborative Filtering (OCCF) mainly deals with the extremely sparse and unbalanced implicit data in one-class problem. Because the data has only a small part of positively labeled feedback, most existing works on OCCF focus on how to model the missing examples [4, 5, 6]. The key idea of these approaches is to introduce weights for treating missing data as negatives. However, most of them use the statistical properties of feedback to determine the weight, which are not very reasonable and not personalized for each user-item pair. As we know, there are rich user and item content information in real-world data, for example, the news content, the article's content and the product's text description, along with the user's profession, hometown and search logs. Although the rating record used by CF brings much improvement to recommendation [7], the enriched content information can assist when we do not have any priori knowledge about the missing data. In the research on One-Class Collaborative Filtering, little has been studied to exploit the content information to help modeling the missing examples of implicit data.

In this paper, we incorporate the user and item content information to assist weighting the unknown data in OCCF. Probabilistic topic modeling method [8] is used to get a content topic feature for each user and item from their content information. Different from the latent factor utilized in matrix factorization CF models [1, 9, 10], the content topic feature is not used to determine the final user-item score, but to determine every unknown user-item pair's weight to be negative. Then we incorporate these weights into the OCCF matrix factorization model and get the final user and item latent factor. The experiments on the data of CiteULike show that the content topic feature can improve performance of the OCCF method.

2 Related Work

Recommender systems are usually classified into three categories [1]: Content-based recommendations [11], Collaborative Filtering (CF) [12] and Hybrid approaches [13].

¹ <http://www.netflixprize.com/>

The Content-based methods make recommendations by analyzing the content of textual information and finding regularities in the content. However, the content information is difficult to collect sometimes and it cannot recommend items which are unseen by user before. Collaborative Filtering methods recommend items to a particular user based on other users with similar patterns of selected items and it does not use the content information. However, it can discover the useful associations between different users and items which Content-based methods ignore. Recently, CF methods based on latent factor models [8, 9, 14] have become very popular. In latent factor models, users and items are represented in a shared latent low-dimensional space of dimension f , that is, user i is associated with a latent vector $u_i \in R^f$ and item j is associated with a latent vector $v_j \in R^f$. Then user i 's rating of item j is approximated by the inner product between their latent factors,

$$\hat{r}_{ij} = u_i^T v_j \tag{1}$$

To get the rating result the major challenge is to compute each item and user's latent factor vectors given an observed matrix of ratings. The common approach is to minimize the regularized squared error on the set of observed ratings,

$$\min_{U,V} \sum_{i,j} (r_{ij} - u_i^T v_j)^2 + \lambda (\|u_i\|_F^2 + \|v_j\|_F^2) \tag{2}$$

where $V=(v_1,v_2,\dots,v_n)$, $U=(u_1,u_2,\dots,u_m)$, $\| \cdot \|_F^2$ is the Frobenius Norm of a matrix and λ is regularization parameter aiming at avoiding overfitting the training data and determined by cross validation. Parameters are often learnt by stochastic gradient descent (SGD) and Alternating least squares (ALS) [9].

However, CF methods are usually faced with the cold start problem when there are few ratings for user or item. Hybrid methods combine Content-based methods and CF to avoid certain limitations of the two methods. However, most of these methods focus on the explicit data, which has obvious difference from the implicit data we want to deal with in this paper.

The most important characteristic of implicit data is that it has no negative examples, which leads to extremely sparsity and unbalance. In prior works, there are several intuitive strategies to handle this problem. One common solution is to treat all the missing data as negative (AMAN), which may bias the recommendation results because many missing data may be positive. Another solution is to treat all the missing data as unknown (AMAU), which ignores the missing ones and only uses the positive ones into the CF models. Instead of these two simple strategies, Pan, et. al [5, 6] proposed a weighted method, which treat all the missing data as negatives, but with weights to respond the confidence to treat them as negative. The loss function is defined as follows:

$$\min_{U,V} \sum_{i,j} w_{ij} (r_{ij} - u_i^T v_j)^2 + \lambda (\|u_i\|_F^2 + \|v_j\|_F^2) \tag{3}$$

A parallel Alternating Least Squares (ALS) method is used to minimize the loss function in Eq. (3) and learn the parameters [19], named wALS. In [5, 6], Pan proposed three weighting schemes: uniform, user-oriented and item-oriented, which mainly use the statistical properties of feedback to determine the weight.

Instead of using a global weighting scheme without any prior knowledge, a better way is to consider the content information of user and item and look at the similarity between them. On the other hand, content analysis based on probabilistic topic modeling has been developed in many researches [10, 15, 16], among which, Latent Dirichlet allocation (LDA) [10] is good at studying latent topic distribution of documents. In this paper, we will use LDA to learn the content topic features from the user and item content information and incorporate them into the weighted OCCF models. Essentially, our approach is a Hybrid method combining weighted latent factor models in OCCF and content analysis based on topic modeling.

3 Our Method

In this section, we first present our problem definition, and then we show how to use the probabilistic topic modeling method to get the content topic feature and incorporate them into the weighted OCCF model. At last, we will summarize our algorithm process briefly.

3.1 Problem Definition

The task in this paper is to find the most interesting recommendations with the implicit feedback data, which is referred to as a top-N recommendation task [17]. Supposing we have m users and n items, the implicit feedback can be expressed by an $m \times n$ matrix R . Taking the CiteUlike data for example, if a user i posted a paper j , the matrix element r_{ij} should be positive and we set it the number 1, naturally we set negative examples as 0. As mentioned before, the implicit data does not have negative examples, thus the matrix R contains only the positive ones originally. Our approach will model the missing examples and decompose the newly modeled matrix R to find the top N sorted items which may be potential positive examples.

3.2 Content Topic Feature Extraction

There usually exists much content information in real-world data. As mentioned in the prior sections, we want to incorporate them into OCCF to assist modeling the missing negative data in implicit feedback. For this purpose, we abstract a content topic feature for each user and item, which describe the latent topic distribution of their content. Latent Dirichlet Allocation (LDA) [10] is a popular probabilistic topic model to discover a set of “topics” from a large collection of documents. It assumes a generative probabilistic model in which documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. We utilize LDA to get the content topic feature in this paper.

In this paper, we consider the CiteULike² “who post what” data as recommended object. The article content information we used here are title and abstract. For each user, the abstract of the article he/she posted before formed a larger article, which is the information for us to learn the user’s content topic feature here.

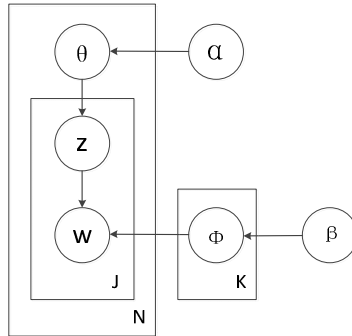


Fig. 1. The graphical model for LDA

Assume there are J words in the vocabulary, K latent topics and N articles. As shown in Figure 1, we denote per-article topic distribution as θ , each a K -dimensional vector, and per-topic word distribution as ϕ , each a J -dimensional vector. The generative process of LDA can be summarized as follows. For each article j in the corpus,

1. Choose topic distribution $\theta_j \sim \text{Dirichlet}(\alpha)$.
2. Choose word distribution $\phi_k \sim \text{Dirichlet}(\beta)$.
3. For each of the word n ,
 - (a) Choose a topic assignment $z_{j,n} \sim \text{Multinomial}(\theta_j)$.
 - (b) Choose a word $w_{j,n} \sim \text{Multinomial}(\phi_{z_{j,n}})$.

For the corpus of articles, we can estimate the parameters θ and ϕ by variational EM [10] or Gibbs sampling method [20]. For each item, θ_j is the content topic feature we want to get, which abstract the topic distribution from the content. Further, for each user, given the user’s content information described before, we can use variational inference to situate its content in terms of the topics and get the content topic feature for user. For simplicity, we note the item content topic feature as q and user content topic feature as p in the rest of the paper.

Similarity between the user and item content topic feature can be measured by the cosine similarity:

$$s(p_i, q_j) = \frac{p_i \cdot q_j}{\|p_i\|_2 \times \|q_j\|_2} \tag{4}$$

² <http://www.citeulike.org/>

3.3 Content Topic Feature Based Weighted OCCF

As mentioned before, Pan, et.al [5, 6] proposed the weighted method in OCCF. He treated all the missing data as negatives and gave them a weight to respond the confidence to treat them as negative. But existing weighting approaches mainly use the statistical properties of feedback to determine the weight, which are not very reasonable and not personalized for each user-item pair. Instead, a better way is to consider the similarity between the user and item content information, that is more similar they are, more likely the user will like the item, thus the less confidence we treat that missing user-item pair as negative. The content topic feature we learned in section 3.2 abstract the topic distribution of the user and item’s content, and it’s reasonable that the content topic feature’s similarity represent the user and item’s content similarity. So we assign the weight to each negative example as:

$$c_{ij} \propto 1-s(p_i, q_j) \tag{5}$$

Notice that we set the weight of each positive example as 1. We define the loss function as follows:

$$Loss(U, V) = \min_{U, V} \sum_{i,j} c_{ij} (r_{ij} - u_i^T v_j)^2 + \lambda (\|u_i\|_F^2 + \|v_j\|_F^2) \tag{6}$$

A parallel Alternating Least Squares method [19], which rotates between fixing user latent factors and item latent factors, is efficient for solving these low rank approximation problems. When all u_i are fixed, the system recomputes the v_j by solving a least squares problem, and vice versa.

Fixing V and solving $\frac{\partial Loss(U, V)}{\partial u_i} = 0$, we have

$$u_i = (VC^{u_i}V^T + \lambda I)^{-1}VC^{u_i}R(u_i) \tag{7}$$

where $V=(v_1, v_2, \dots, v_n) \in \mathbb{R}^{f \times n}$, C^{u_i} is a $n \times n$ diagonal matrix with the elements $C_{jj}^{u_i} = c_{ij}$, I is an $f \times f$ identity matrix, and $R(u_i)$ is a n dimensional vector that contains the ratings by u_i for all the item. Similarly, given fixed U, we can recompute V as

$$v_j = (UC^{v_j}U^T + \lambda I)^{-1}UC^{v_j}R(v_j) \tag{8}$$

where $U=(u_1, u_2, \dots, u_m) \in \mathbb{R}^{f \times m}$, C^{v_j} is a $m \times m$ diagonal matrix with the elements $C_{ii}^{v_j} = c_{ij}$ and $R(v_j)$ is a m dimensional vector that contains all the users’ ratings for item v_j . We repeat these iterative update until convergence to get the final user and item latent factors. For a particular user i , we will recommend the N items with the largest value of $u_i^T v_j$.

As discussed before, the implicit data is unbalanced with no negative feedback. If we treat all the missing data as negative, it's still unbalanced with too many negative examples, which is costly to learn the user and item latent factor and may bias the result. Thus, we sample negative examples from missing data before weighted OCCF. The sampled probability for a missing user-item pair is also based on the content topic feature similarity:

$$P_{ij} \propto 1/s(p_i, q_j) \quad (9)$$

That is, the more similar between the user and item content information, the lower probability we should sample them as negative. Based on this sampling scheme, we sample comparable number of negative examples to positive examples.

To summarize our algorithm process, the detailed process can be depicted by Figure 2.

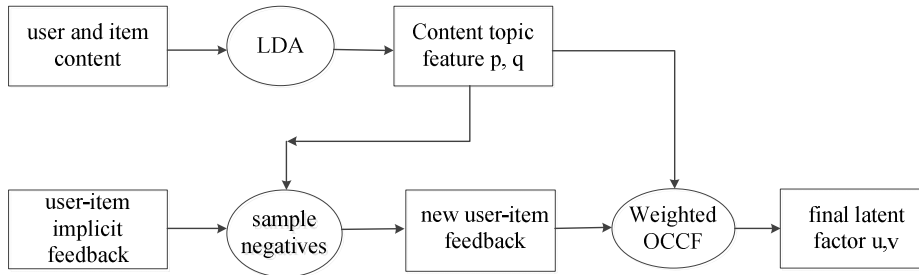


Fig. 2. The flowchart of our algorithm

From the flowchart of the proposed algorithm, our method incorporates the content information by probabilistic topic model LDA to assist modeling the missing negative examples in implicit feedback. When there is no prior knowledge, the content information may help us to decide the negatives. We use CTF-wOCCF (Content Topic Feature based weighted OCCF) to denote our method in the following sections.

4 Experiments

To evaluate the proposed method, we conducted experiment on a subset of CiteUlike³ “who post what” data. CiteUlike is a scientific article sharing service where users create personal libraries by posting the articles they like. Each article has information such as title, abstract, authors, publications and keywords. As mentioned before, the content information we used contains the title and abstract. The subset we used contains 5551 users and 16980 articles with 204,986 observed user-item pairs, which is very sparse. In this subset, every user has at least 10 articles posted in the library, 93% of the users have no more than 100 articles, and 97% of the articles appear in fewer than 40 libraries.

³ <http://www.citeulike.org/faq/data.adp>

In our experiment we randomly divide the datasets into training and testing sets in the ratio of 80% to 20% and repeat the random splits 20 times to get the average result.

We compared our approach with the following baselines. Most of the baselines have been discussed in section 2. We present them here briefly.

1. AMAN: treating all missing data as negative. Solving the equation 2 with all the positive examples assigned to 1 and all the missing examples assigned to 0. Alternating least squares (ALS) optimization procedure is adopted to solve the factorization.
2. AMAU: treating all missing data as unknown. Solving the equation 2 with all the positive examples assigned to 1 and ignoring the missing ones. Stochastic gradient descent optimization procedure is adopted to solve the factorization.
3. wALS (uniform): Weighted method in OCCF described in section 2. The weight is uniform with value less than 1 [6, 5].
4. wALS (item-oriented): The weights are not uniform and the j^{th} item has the weight proportional to $m - \sum_i r_{ij}$, which takes the intuition that if an item is viewed by less users, the missing data for this item is negative with higher probability [6, 5].
5. wALS (user-oriented): The weights are not uniform and the i^{th} user has the weight proportional to $\sum_j r_{ij}$, which takes the intuition that if a user has viewed more items, those items that he has not viewed could be negative with higher probability [6, 5].
6. LDA: Here, LDA represent a content-based model that only uses LDA-content topic features as we discussed in section 3.2. The ratings are computed by the similarity of user and item content topic features p and q .

4.1 Evaluation Criteria

As top-N recommendation task, we present each user with N articles sorted by their predicted rating and evaluate based on the articles actually posted in the user's library. Our testing methodology is similar to the one described in [17], after training the model over training data, we compute the predicted ratings over a probe set to find the top N articles of the probe set. This probe set is obtained by randomly selecting 2000 additional articles unrated for each user and adding the test set. For each training-testing splitting, the probe set is sampled 20 times and the results reported are averaged over them.

As described in [17], to evaluate our top N results, we assume that most of the unrated items in probe set are not interesting to users, thus the Precision and Recall are computed as follows,

$$Precision(N) = \frac{\#TestsetHits(N)}{N} \quad (10)$$

$$Recall(N) = \frac{\#TestsetHits(N)}{\#TestsetPositives} \quad (11)$$

where N is the number of recommended items, $\#TestsetHits(N)$ is the number of items contained in test set in top- N recommended results, $\#TestsetPositives$ is the total number of items the user likes in the test set.

MAP (Mean Average Precision) assesses the overall performance based on precisions at different recall levels. It calculates the mean of average precision (AP) over all users in the test set. AP for user u is the average of precisions computed at all positions with a preferred item:

$$AP_u = \frac{\sum_{i=1}^N Precision(i) \times pref(i)}{\# \text{ of preferred items}} \quad (12)$$

where $Precision(i)$ is the precision at ranked position i , $pref(i)$ is a binary indicator returning 1 if the i -th item is preferred or 0 otherwise.

4.2 Experimental Results

Table 1 show the Precision, Recall and MAP results when return 20 top articles for each of the 6 baselines, and compare them with our method CTF-wOCCF (Content Topic Feature based weighted OCCF). We try different number of final latent factor dimension f from 50 to 200, and find that the performance increases with the f increasing. $f=200$ is set in reporting the results. For CTF-wOCCF we set the content topic feature dimension $k=50$ while the final latent factor dimension is the same as other baselines.

Table 1. Comparison of all methods in terms of MAP, Recall and Precision

Methods	MAP	Recall	Precision
AMAU	0.0161	0.0417	0.0185
LDA	0.1193	0.3659	0.1099
AMAN	0.1735	0.3943	0.1350
wALS(uniform)	0.2324	0.4563	0.1358
wALS(item)	0.2129	0.4347	0.1305
wALS(user)	0.2114	0.4372	0.1306
CTF-wOCCF	0.2840	0.5399	0.1630

As expected, since it only uses a small number of positive examples, AMAU returns the worst performance. Even the content-based method LDA outperforms AMAU, which shows that the content information helps much in article recommendation. AMAN performs better than AMAU and LDA, shows that negative

examples are necessary for recommendation with implicit data. The three weighted methods all increase the result of AMAN, demonstrates that it's useful to model the negative examples in implicit data. However, the user-oriented and item-oriented weighting schemes do not perform as well as simple uniform scheme. The reason may be that our article posting data is very sparse with most users have few articles and most articles appear in few libraries, which leads that weighting scheme based on the number of ratings for each user or item become very weak. However, compare to the best baseline (uniform wALS) results, our method leads to significant improvements of 0.0516 in MAP, which suggests that the content information is useful in article recommendation and our method to incorporate the useful content information into weighted OCCF by content topic feature is suitable for implicit data.

In Figure 3, we show the MAP of all methods in different number of recommended articles. The results demonstrate that our method is better than the compared algorithms varying with the number of recommended articles. Thus the conclusion we get from Figure 3 is consistent with that of table 1.

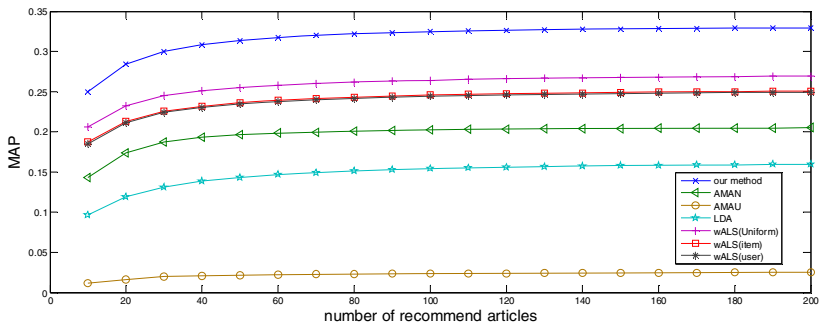


Fig. 3. Performance varying the number of recommended articles

5 Conclusions

In this paper, we proposed a novel weighted one-class collaborative filtering (OCCF) method. The proposed method integrates the useful content information into OCCF to deal with the one class problem in implicit data based recommendation. To incorporate the content information, we learn a content topic feature for each user and item by probabilistic topic model LDA to assist modeling the missing negative examples in the weighted one-class collaborative filtering. Experiment on CiteUlike data shows that our method outperforms state of the art algorithms, which suggests that the content information is useful to overcome the sparsity and unbalance in OCCF and our method to incorporate the useful content information into weighted OCCF by content topic feature is effective.

Acknowledgments. This work was supported in part by the 973 Program under Project 2010CB327905, by the National Natural Science Foundation of China under Grant nos. 61170127, 60975010, 60833006, and 61070104.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. on Knowl. and Data Eng.* 17(6), 734–749 (2005)
2. Linden, G., Smith, B., York, J.: Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing* 7, 76–80 (2003)
3. Sarwar, B., Karypis, G., Konstan, J., Riedl, J.: Item-based collaborative filtering recommendation algorithms. In: *International Conference on the World Wide Web*, pp. 285–295 (2001)
4. Hu, Y., Koren, Y., Volinsky, C.: Collaborative filtering for implicit feedback datasets. In: *IEEE International Conference on Data Mining*, pp. 263–272 (2008)
5. Pan, R., Scholz, M.: Mind the gaps: weighting the unknown in large-scale one-class collaborative filtering. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 667–676 (2009)
6. Pan, R., Zhou, Y., Cao, B., Liu, N.N., Lukose, R., Scholz, M., Yang, Q.: One-class collaborative filtering. In: *IEEE International Conference on Data Mining*, pp. 502–511 (2008)
7. Pilaszy, I., Tikk, D.: Recommending new movies: Even a few ratings are more valuable than metadata. In: *Proceedings of the Third ACM Conference on Recommender Systems*, pp. 93–100 (2009)
8. Salakhutdinov, R., Mnih, A.: Probabilistic matrix factorization. In: *NIPS* (2007)
9. Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. *Computer* 42(8), 30–37 (2009)
10. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. *Journal of Machine Learning Research* 3, 993–1022 (2003)
11. Balabanovic, M., Shoham, Y.: Fab: Content-based, collaborative recommendation. *Communications of the ACM* 40(3), 66–72 (1997)
12. Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. *Adv. In: Artif. Intell.* (2009)
13. Claypool, M., Gokhale, A., Miranda, T., Murnikov, P., Netes, D., Sartin, M.: Combining content-based and collaborative filters in an online newspaper. In: *Proceedings of ACM SIGIR Workshop on Recommender Systems* (1999)
14. Agarwal, D., Chen, B.-C.: Regression-based latent factor models. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 19–28 (2009)
15. The, Y., Jordan, M., Beal, M., Blei, D.: Hierarchical Dirichlet processes. *Journal of the American Statistical Association* 101(476), 1566–1581 (2007)
16. Hofmann, T.: Latent semantic models for collaborative filtering. *ACM Trans. Inf. Syst.* 22(1), 89–115 (2004)
17. Cremonesi, P., Koren, Y., Turrin, R.: Performance of recommender algorithms on top-N recommendation tasks. In: *Proceedings of the Fourth ACM Conference on Recommender Systems*, pp. 39–46 (2010)
18. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: *SIGIR 1999*, pp. 230–237 (1999)
19. Zhou, Y., Wilkinson, D., Schreiber, R., Pan, R.: Large-Scale Parallel Collaborative Filtering for the Netflix Prize. In: *Fleischer, R., Xu, J. (eds.) AAIM 2008. LNCS*, vol. 5034, pp. 337–348. Springer, Heidelberg (2008)
20. Griffiths, T.L., Steyvers, M.: Finding scientific topics. *Proceedings of the National Academy Science* 101(suppl.1), 5228–5235 (2004)

Contextualizing Tag Ranking and Saliency Detection for Social Images

Wen Wang, Congyan Lang, and Songhe Feng

School of Computer & Information Technology,
Beijing Jiaotong University, Beijing, 100044, China
{11120483,cylang,shfeng}@bjtu.edu.cn

Abstract. Tag ranking and saliency detection are two key tasks for image understanding, and have attracted much attention in the past decades. In this paper, we investigate how to iteratively and mutually boost tag ranking and saliency detection by taking the outputs from one task as the context of the other one. Our method first computes an initial saliency value based on fusing multiple feature maps, and then iteratively refines saliency map based on the contextual information from image tag ranking. As a result, an integrated framework for tag saliency ranking which combines both visual attention model and multi-instance learning to investigate the saliency ranking order information. We show that this mutual reinforcement of saliency detection and tag ranking improves the performance by using this combined approach. Experiments conducted on Corel and Flickr image datasets demonstrate the effectiveness of the proposed framework.

1 Introduction

The tag ranking task aims to rank tags according to their semantic relevance with respect of the given image, whereas the goal of saliency detection is to find the image areas where one or more of their features differ from those in the surroundings. Traditionally, these two tasks are considered individually.

Visual saliency has been extensively studied in signal processing, computer vision, machine learning, psychology and vision research literatures (e.g., [4,2,3,8,5,6]). The traditional paradigm that determines saliency by only using the visual properties of the image. These methods suffer from a limitation in that cluttered background may produce higher saliency because such backgrounds possess high global energy in the cases of complex scenes. Meanwhile, object borders are often assigned higher saliency than the salient regions and thus they may not produce satisfactory results. The rapid popularization of digital cameras and mobile phone cameras has led to an explosive growth of social image sharing websites, such as Flickr. Such social images are annotated with tags[14]. As shown in Figure1, from the perspective of human perception, one can observe that the tag building is apparently salient than the tag sky. And people tend to pay more attention to the region with "building" tag than the others. Intuitively, the tag can be used as outside-image context for saliency detection. While previous methods usually define saliency as image regions whose features differ from those in the image, we treat the saliency of the image as the regions

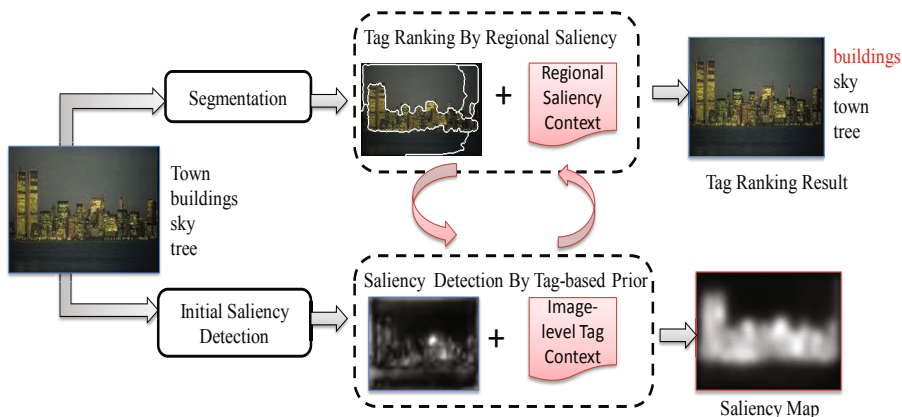


Fig. 1. The flowchart of our contextual method for tag ranking and saliency detection. We propose a novel method to iteratively and mutually boost tag ranking and saliency detection by taking the outputs from one task as the context of the other one.

differ from those in the image and correspond to visual content described by the first tag.

Social images are annotated with orderless tags, which limited the effectiveness of image search and retrieval applications[15,16,20]. So automatic image tag ranking which aims to rank tags according to their semantic relevance with respect of the given image, has attracted a lot of attention [9,19]. Existing tag ranking methods focused on ranking the tags from the relevance aspect. Li et al. [11] introduced an approach that learns the relevance scores of tags by a neighborhood voting method. Given an image and one of its associated tags, the relevance score is learned by accumulating the votes from the visual neighbors of the image. More recently, Liu et al. [10] utilized the Kernel Density Estimation (KDE) to estimate the relevance score of each tag individually, and further performed a random-walk based refinement to boost tag ranking performance by exploring the relationship of tags. In general, users often pay much attention to the first one or two tags that annotated on the image, since these tags might describe the main content of the image and can assist users easily manage and access large-scale image dataset. Intuitively, a tag is visually salient (representative) if all the images annotated with the tag are visually similar to each other. The essential idea behind the tag ranking is to find the most representative images with respect to a given tag. In contrast to the tag relevance ranking algorithm, we have proposed the tag saliency ranking algorithm [12], which is able to rank tags according to their saliency values to an images content. The tag saliency ranking algorithm firstly locates tags to the corresponding regions with a multiple-instance learning approach, and then analyzes the saliency values of these regions. It can provide more comprehensive information when an image is relevant to multiple tags. As to the tag saliency ranking, the main disadvantage lies in that such algorithm heavily relies on the saliency map constructed by the visual attention model.

In this paper, we firstly propose a new paradigm on saliency detection, which aims at producing more reliable results by mining the context information from ranked tags, namely modeling the cross-media information from images with ranked tags. Then, we propose an iterative framework such that the performance of saliency detection and tag ranking can be iteratively and mutually boosted as shown in Figure 1.

2 Tag Saliency Ranking Combining MIL with Visual Attention Model

Firstly, we adopt our tag saliency ranking strategy previously proposed in [12]. Given an image and its associated tags, we first estimate the relevance between each tag and segmented region individually through the multi-instance learning (MIL) algorithm. Then each segmented region can be assigned a saliency value to reflect the importance in the given image. Finally, the tags of the image can be ranked according to the corresponding regions saliency value. We utilize a reinforced DD algorithm [18] to accomplish the label propagation from image-level to region-level. Specifically, the concept of instance prototypes (IPs) is proposed to denote the semantic concept instead of one single target concept.

Given a specific tag $w \in V$, we denote a certain positive bag for tag w as B_i^{w+} and its j -th instance (region) as r_{ij}^{w+} ($j = 1, \dots, n_i^{w+}$). The negative bag and the corresponding instance are defined similarly as B_i^{w-}, r_{ij}^{w-} respectively. We denote the total training set $L = L^{w+} \cup L^{w-}$ according to whether the tag w is associated with the image. Given the training set $L = L^{w+} \cup L^{w-}$, for each bag $B_i^{w+} \in L^{w+}$ ($i = 1, \dots, l^{w+}$), the DD value of each instance $r_{ij}^{w+} \in B_i^{w+}$ ($j = 1, \dots, n_i^{w+}$) is defined by:

$$DD(r_{ij}^{w+}, L) = \sum_{m=1}^{|L|} \max_n \{1 - |y_m - \exp(-dist^2(B_{mm}, r_{ij}^{w+}))|\} \quad (1)$$

Once the DD values of all the instances are computed, all the instances satisfying the DD values larger than the pre-defined threshold are selected as instance prototypes for each tag w , Gaussian mixture model (GMM) is employed here to learn the class-conditional density $P(x|w)$ for w . Then, a feature mapping strategy is developed to measure the likelihood that these Gaussian components are present in each positive bag. For each instance $r_{ij}^{w+} \in B_i^{w+}$ with respect to the tag w , the probability that region r_{ij}^{w+} is assigned to the given tag w is defined as:

$$P(w|r_{ij}^{w+}) = \sum_{i=1}^M \pi_i P(r_{ij}^{w+}|w) \quad (2)$$

Once given the image $I = \{r_1^I, \dots, r_m^I\}$ and the associated tag list $W = \{w_1^I, \dots, w_n^I\}$, we can utilize the aforementioned algorithm to perform label-to-region task. Each segmented region r_i^I ($i = 1, \dots, m$) will be assigned a unique tag w_j^I ($j = 1, \dots, n$) according to the probability value that it belongs to. And finally, tags can be ranked by the averaged saliency scores of the corresponding regions in the given image in descending order.

3 Saliency Detection with Image-Level Ranked Tag Priors

We seek the context information from the ranked tags in order to improve the performance of the initial bottom up saliency detection models. In this section, we propose to model the relationship between tag and saliency by approximating the joint density with a Mixture of Gaussians.

For the image $I = \{r_1^I, \dots, r_m^I\}$ and the associated ranked tag list $W = \{w_1^I, \dots, w_n^I\}$, we can interpret the saliency in terms of different mechanisms that contribute to the guidance of attention as:

$$S(x) = P(s|x)P(s_{w^*}|x)P(s_{w^*}|w^*, \ell) \quad (3)$$

For each pixel $x \in I$, the first term does not depend on the tag context, and therefore is a pure bottom-up factor. This term fits the definition of saliency [1]. And the second term $P(s_w|x)$ represents the tag saliency. The hypothesis underlying is that the regions corresponding to the first tag are considered more informative and therefore will attract attention. We denote the tag saliency as $P(s_{w^*}|x) = P(w^*|r_i^I), x \in r_i^I$. Besides the tag saliency, we model the relationship between first tag and saliency by approximating the joint density with a mixture of gaussians and we seek the global-context priors in order to improve the performance of the saliency detection. Hence, the third term can be denoted as global-context priors by mining from the image subset for the tag w^* . The role of this term in the model is to activate the locations most likely to contain the region with the tag w^* and reducing the saliency of image regions not relevant for the tag.

Formally, let ℓ and S represent the location of the region with tag w^* and fixation map of the image I , respectively. For the vector s_{w^*} , larger (smaller) magnitude implies that the pixel is more salient (less salient). The joint density between saliency response and location distribution is written as

$$p(s_{w^*}, \ell|w^*) = \sum_{k=1}^K P(k)p(s_{w^*}|w^*, k)P(\ell|w^*, k), \quad (4)$$

where k indicates the k th component of the GMM. From the joint distribution we calculate the conditional density required for the location modulated saliency:

$$\begin{aligned} p(s_{w^*}|\ell, w^*) &= \frac{P(s_{w^*}, \ell)}{\sum_{k=1}^K P(k)P(\ell|w^*, k)} \\ &= \sum_{k=1}^K \pi_k \mathcal{N}(s_{w^*}; \mu_k, \Lambda_k) \mathcal{N}(\ell; \nu_k, \Upsilon_k), \end{aligned} \quad (5)$$

The parameters of the model are obtained from the training dataset and the EM algorithm is applied for fitting Gaussian mixtures. For the image I_t , its corresponding saliency map can be defined as the predicted saliency $s_{w^*}^t$.

4 Experiments and Results

4.1 Experiment Setup

To evaluate the effectiveness of the proposed method, we tested the performance on COREL and NUS-WIDE-Object dataset. The vocabulary of COREL contains 371 different keywords and each image is associated with 4-5 keywords. The NUS-WIDE-Object Flickr dataset [17] includes 30,000 images and their associated tags. For images which do not contain any distinct object, the tag saliency ranking algorithm may not get satisfactory results. Two students are selected to manually label each image in the Corel5K and NUS-WIDE-Object datasets as attentive or non-attentive. We conduct our experiment on the attentive image set. Since the collected tags are rather noisy, we keep tags that belong to the noun-tags and tags appear with too low frequencies are filtered out. Finally around 500 tags are kept after these processes. Images with no tag are not used and thus we obtain a subset with 6,850 images, and in average 4 tags associated with each image.

Images in both datasets are segmented into image regions using JSEG algorithm, and the 64-D color histogram and 73-D edge direction histogram are extracted for both global and region level. As to the ground-truth generation, we observe that the ranking order of each tag-list annotated on the image in the Corel5K dataset can well fit human perception, so we just use the existing tag ranking orders released along with the dataset as the ground truth for comparison. As to the NUS-WIDE-Object Flickr dataset, three students are selected as volunteers to rank the tag list from the perspective of human perception. The voting results are considered as the ground truth. All approaches in the experiments are executed on a PC with Intel 3.0GHz CPU and 4G memory.

The performance is evaluated by average precision (AP) and mean average precision (MAP), which are commonly adopted in the evaluations. Average precision measures ranking quality of the whole list. Since it is an approximation of the area under the precision-recall curve, AP is commonly considered as a good combination of precision and recall. To calculate AP for one image tag ranking result, given an arbitrary image I in the dataset with the ranked tag list $T = \{t_1, t_2, \dots, t_N\}$ based on the proposed method, the average precision of the ranking result as compared with the ground-truth tag list is defined as

$$AP(I, T) = \frac{1}{N} \sum_{i=1}^N \frac{R_i \delta(i)}{i},$$

where N is the length of the tag list annotated on the given image, R_i the number of corresponding correct tags in the top i ranked tag list. Here $\delta(*)$ is an indicate function, i.e., $\delta(*)$ returns a value of 1 if i -th tag ranked by the proposed method equals to the ground-truth in the same position and 0 otherwise. To evaluate the overall performance, we use mean average precision (MAP) which is the mean value of the AP over all images in the tag ranking experiment.

4.2 Experimental Results

We compare our proposed tag ranking method with the following methods in terms of MAP. One is purely tag relevance ranking via k NN based (k NN-NV).



Fig. 2. Examples of tag ranking results based on the proposed method

Table 1. Precision comparison with different algorithms on COREL & NUS-WIDE-Object dataset

Algorithms	Precision(Core15K)	Precision(NUS-WIDE)
KNN-NV	55.32%	50.24%
TSR	65.73%	59.27%
The Proposed	68.65%	62.13%

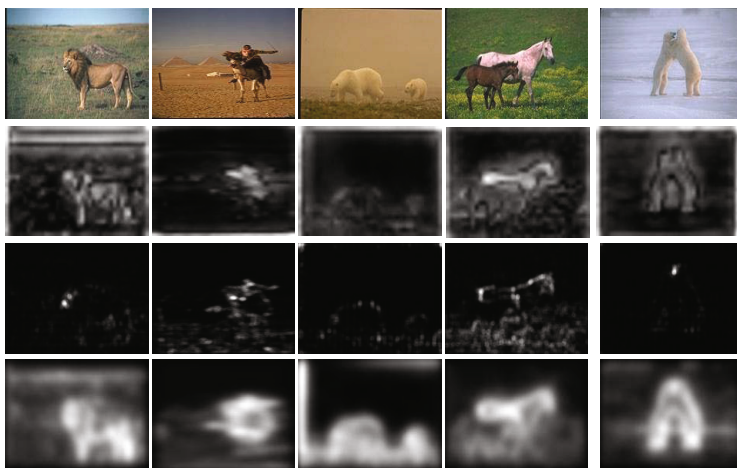


Fig. 3. Visual comparison of saliency maps. From top to bottom: the input image, saliency map generated by method GBVS, SR, and our method, respectively

And the other one is tag saliency ranking method (abbreviated as TSR) only using initial saliency map. We choose k equals to 100 in our experiment, i.e., 200 nearest images are used to learn tag ranking by kNN-NV. Our method can be deemed as an iterative and mutual framework to boost tag ranking and saliency detection by taking the outputs from one as the context of the other one. In this experiment, the mutual contextualization is conducted for 3 iterations.

Figure 2 illustrates several exemplary ranking results, from which we can clearly see that the tag ranking lists are better than the original ones. From the figure, we can see that the tag saliency ranking strategy is utilized and tags corresponding to the salient part are emphasized. The performance comparison is shown in Table 1. From the results, we can conclude that, since the visual neighbor voting technique is utilized on a relatively small image dataset, the ranking performance is not very satisfying as compared with our method.

Furthermore, we present comparisons with our saliency detection method against the two leading methods in the field of saliency map computation. One is the classical center-surround based salient model presented by Harel [3], denoted as GBVS. Another approach is Hous method [7], which is based on the spectral residual analysis, and we call it “SR”. Figure 3 gives some of the predicted saliency maps. It can be seen that our proposed method exhibits stronger consistence with human eye fixations. These results validate with confidence the effectiveness of our proposed method, which integrates both the image features and context information from ranked tags into inference procedure.

5 Conclusions and Future Work

In this paper, we introduce an iterative and mutual framework to boost tag ranking and saliency detection by taking the outputs from one as the context of the other one for social images. Firstly, the tag saliency ranking approach is proposed by emphasizing distinct objects in the image. On the other hand, a new paradigm is presented on saliency detection, which aims at producing more reliable results by mining the context information from ranked tags. In future work, we are interested in how to integrate two tasks into a joint inference procedure instead of the current iterative model.

Acknowledgements. This work is partially supported by National Nature Science Foundation of China (61272352, 61100142, 60972145); Beijing Jiaotong University Science Foundation (No. 2011JBM218, 2011JBM219); and the Doctoral Fund of Ministry of Education of China(20110009120005).

References

1. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. TPAMI (1998)
2. Bruce, N., Tsotsos, J.: Saliency based on information maximization. In: NIPS (2006)
3. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. In: NIPS (2006)
4. Tsotsos, J., Culhane, S., Wai, W., Lai, Y., Davis, N.: Modelling visual attention via selective tuning. Artificial Intelligence (1995)
5. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV (2009)
6. Meur, O., Chevet, J.: Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks. TIP (2010)
7. Hou, X., Zhang, L.: Dynamic visual attention: searching for coding length increments. In: NIPS (2008)

8. Lang, C., Liu, G., Yu, J., Yan, S.: Saliency detection by multi-task sparsity pursuit. *TIP* (2011)
9. Wang, M., Ni, B., Hua, X., Chua, T.: Assistive Tagging: A Survey of Multimedia Tagging with Human-Computer Joint Exploration. *ACM Computing Survey* (2012)
10. Liu, D., Hua, X.S., Yang, L.J., Wang, M., Zhang, H.J.: Tag Ranking In: *WWW* (2009)
11. Li, X.R., Snoek, C.G.M., Worring, M.: Learning Social Tag Relevance by Neighbor Voting. *IEEE Trans. on Multimedia* (2009)
12. Feng, S., Lang, C., Xu, D.: Beyond tag relevance: integrating visual attention model and multi-instance learning for tag saliency ranking In: *ACM CIVR* (2010)
13. Wang, C.H., Yan, S.C.: Multi-Label Sparse Coding for Automatic Image Annotation In: *IEEE CVPR* (2009)
14. Gao, S., Wang, Z., Chia, L.: Automatic Image Tagging via Category Label and Web Data In: *ACM Multimedia* (2010)
15. Tang, J., Hong, R., Yan, S., Chua, T.: Image Annotation by kNN-Sparse Graph-based Label Propagation over Noisily-Tagged Web Images. *ACM Trans. on Intelligent Systems and Technology* (2011)
16. Tang, J., Yan, S., Hong, R., Qi, G., Seng Chua, T.: Inferring semantic concepts from community-contributed images and noisy tags. In: *ACM Multimedia* (2009)
17. Chua, T., Tang, J., Hong, R.: NUS-WIDE: A Real-World Web Image Database from National University of Singapore In: *ACM CIVR* (2009)
18. Rahmani, R., Goldman, S.: MISSL: multiple-instance semi-supervised learning In: *ICML* (2006)
19. Wang, M., Yang, K., Hua, X., Zhang, H.: Towards a Relevant and Diverse Search of Social Images. *IEEE Trans. on Multimedia* (2010)
20. Wang, M., Hong, R., Li, G., Zha, Z.: Event Driven Web Video Summarization by Tag Localization and Key-Shot Identification. *IEEE Transactions on Multimedia* (2012)

Illumination Variation Dictionary Designing for Single-Sample Face Recognition via Sparse Representation

Biao Wang^{1,2,3}, Weifeng Li^{1,2,3}, and Qingmin Liao^{1,2,3}

¹ Department of Electronic Engineering/Graduate School at Shenzhen,
Tsinghua University, China

² Tsinghua-PolyU Biometric Joint Laboratory, Tsinghua University, China

³ Shenzhen Key Laboratory of Information Science and Technology, China
wangbiao08@mails.thu.edu.cn, li.weifeng@sz.tsinghua.edu.cn,
liaoqm@tsinghua.edu.cn

Abstract. This paper focuses on enhancing Sparse Representation based Classifier (SRC) in single-sample face recognition tasks under varying illumination conditions. The major contribution is two-fold: firstly, we present an interesting observation based on Lambertian reflectance model: the identity information will be canceled out by the pair-wise difference images from the same subject in logarithmic domain, and only the subject-independent illumination variation retains. Secondly, inspired from this observation, we propose to “borrow” illumination variations from any generic subject by constructing an illumination variation dictionary composed of pair-wise difference images of generic subjects in logarithmic domain to cover the possible illumination variations between test and gallery samples. Experimental results on Extended Yale B and FERET face databases demonstrate the superiority of our method.

Keywords: Face recognition, single-sample problem, sparse representation, illumination variation dictionary.

1 Introduction

During the last two decades, face recognition remains a very active topic in computer vision communities. Although a lot of effective methods have been proposed [1] [2], robust face recognition under variant illumination conditions is still challenging [3], especially when there is only a single sample per subject [4], which is common in many face recognition applications, e.g., law enforcement and access control, either due to the laborious work to collect multiple samples or the heavy cost to store and process them.

One of the most exciting breakthroughs for face recognition in recent years is the Sparse Representation based Classifier [5]. However, robust face recognition via SRC needs a large set of training samples for each subject to span the variant variations (e.g., illumination, expression) for the test sample of that subject. Wagner *et al.* proposed to acquire multiple samples for each subject under variant

illumination conditions [6]. Experimental results have shown SRC degrades a lot when there is only single sample per subject [6]. To improve the robustness of SRC in single sample face recognition applications, Chang *et al.* [7] propose to enlarge the training samples by generating virtual samples via image shifting and PCA-based reconstructing. In a similar manner, Wang *et al.* [8] take advantage of geometric transformation and svd decomposition for virtual sample generation. Generally speaking, these methods need prior knowledge for the generation of virtual samples. However, it's still an open problem to guarantee the quality and reality of the generated virtual samples.

Very recently, Extended SRC (ESRC) proposed by Deng *et al.* [9] takes advantage of the pair-wise difference images from a generic database (outside the gallery) consisting of multiple generic subjects, each with multiple samples with variant variations, to construct the intra-class variation dictionary, which is further incorporated into the framework of SRC to cover the variations between gallery and test samples. The basic idea of ESRC is that the intra-class variation of each subject can be “borrowed” from a sparse linear combination of the intra-class difference from sufficient number of generic subjects, in which it's assumed that the faces with similar shape to the test subject could be found. Therefore, towards large-scale face recognition applications, ESRC requires a generic database with relatively large size. However, building such a generic database is not only laborious but also will result in heavy computation cost (l^1 -minimization solver) in the recognition stage.

In this paper, we lift the aforementioned assumption and propose to construct the illumination variation dictionary by pair-wise difference images in logarithmic domain rather than those in pixel domain, with whose help we can extend ESRC from “borrowing illumination variation from similar generic subjects” to “borrowing illumination variation from any generic subjects”, thus make it much easier to build a suitable generic database. This is motivated by the observation based on the Lambertian reflectance model: the identity information of a specific subject will be canceled out by the pair-wise difference images in logarithmic domain, which implies that the illumination variation of each subject can be approximated by a sparse linear combination of the illumination variation from a generic database with relatively small size, resulting in the reduction of both the computation cost to solve the sparse solution and the laborious work in constructing large-scale generic database. The experimental results show that incorporating the proposed illumination variation dictionary designing scheme into the ESRC framework can largely enhance the performance of single-sample face recognition tasks under varying illumination conditions. Particularly, compared with original ESRC proposed in [9], with only one generic subject, our proposed method notably improves the recognition rates on Subset 4 and 5 of Extended Yale B dataset, which are under harsh illumination conditions, from 49.28%, 8.54% to 84.06%, 71.89%, respectively.

2 Illumination Variation Dictionary

For single-sample face recognition, we denote the gallery samples as $A = [u_1, u_2, \dots, u_c] \in \mathbb{R}^{d \times c}$, in which $d = w \times h$ is the image dimension, and c is the number of subjects, and u_i is the gallery sample for subject i . A new test sample $y \in \mathbb{R}^d$ belongs to subject i can be expressed as:

$$y = y_0 + e_0 = Ax_0 + e_0, \quad (1)$$

where $e_0 \in \mathbb{R}^d$ is the intra-class variations between gallery and test samples, and $x_0 \in \mathbb{R}^c$ is a sparse vector whose nonzero entries are only associated with subject i .

We are merely concerned the illumination variations in this paper. Typically, in face recognition, the gallery image is usually normally illuminated, while the test image may be arbitrarily illuminated. The illumination cone theory suggests that a wide range of illuminations can be represented by a smaller number of illumination [10]. Therefore it's reasonable to assume that the illumination variation e_0 has a sparse representation with respect to some dictionary $A_e \in \mathbb{R}^{d \times n_e}$, which we call the *illumination variation dictionary*. Eq. (1) could be re-expressed as:

$$y = y_0 + e_0 = Ax_0 + A_e \gamma_0. \quad (2)$$

Compressive sensing theory implies that if the solution is sparse enough, it could be recovered by solving the following l^1 -minimization problem:

$$\begin{aligned} \begin{bmatrix} \hat{x}_0 \\ \hat{\gamma}_0 \end{bmatrix} &= \arg \min \left\| \begin{bmatrix} x \\ \gamma \end{bmatrix} \right\|_1, \\ \text{s.t.} \quad &\left\| \begin{bmatrix} A & A_e \end{bmatrix} \begin{bmatrix} x \\ \gamma \end{bmatrix} - y \right\|_2 \leq \varepsilon. \end{aligned} \quad (3)$$

The identity of y could be determined by:

$$\text{Identity}(y) = \arg \min_i \left\| y - \begin{bmatrix} A & A_e \end{bmatrix} \begin{bmatrix} \delta_i(\hat{x}_0) \\ \hat{\gamma}_0 \end{bmatrix} \right\|_2, \quad (4)$$

where $i = 1, \dots, c$, and $\delta_i(\hat{x}_0) \in \mathbb{R}^c$ is a vector whose only nonzero entries in \hat{x}_0 are associated with subject i .

Deng *et al.* [9] proposed ESRC to utilize the pair-wise difference images from a generic face database with multiple samples per subject to construct the illumination variation dictionary:

$$\begin{aligned} A_e &= [D_1, D_2, \dots, D_{N_g}], \\ D_i &= [v_{i,1} - v_{i,0}, v_{i,2} - v_{i,0}, \dots, v_{i,N_g^i-1} - v_{i,0}], \end{aligned} \quad (5)$$

in which N_g is the total number of generic subjects, and N_g^i is the sample number of generic subject i . $v_{i,j}, j \geq 0$ refers to the j -th sample for generic subject i , and $v_{i,0}$ usually refers to the sample image for generic subject i with normal illumination conditions. The ESRC can be used for single-sample face recognition, as it can recover the variation of a test sample by ‘‘borrowing’’ the image differences from the subjects in a generic database.

3 Proposed Illumination Variation Dictionary

3.1 Analysis on ESRC via Lambertian Reflectance Model

Lambertian reflectance model suggests that a face image F can be expressed by

$$F(x, y) = R(x, y)I(x, y), \quad (6)$$

in which $F(x, y)$ is the image pixel value, and $R(x, y)$ is the reflectance, and $I(x, y)$ is the illuminance at each pixel (x, y) . I depends on the lighting source, while R depends only on the characteristics of the facial surface, which includes both the albedo (surface texture) and the surface normal (3D shape), and thus can be regarded as the intrinsic part for human identity.

For test and gallery sample belong to subject i , assuming the only variation between the two samples are caused by illumination condition, we could denote them as $F_{i,t} = R_i I_t$, and $F_{i,g} = R_i I_g$, respectively. In ESRC [9], to sparsely represent $F_{i,t}$ correctly, the nonzero coefficients should lie on $F_{i,g}$ and the atoms in A_e corresponding to the following variations:

$$F_{i,t} - F_{i,g} = R_i I_t - R_i I_g = R_i (I_t - I_g). \quad (7)$$

For a generic subject p , the difference image for two samples under two illumination conditions, denoted as I_1 and I_0 , could be represented by

$$F_{p,1} - F_{p,0} = R_p I_1 - R_p I_0 = R_p (I_1 - I_0). \quad (8)$$

As seen from Eq. (7) and Eq. (8), the identity information, denoted as R_i or R_p , still exists in the difference image. In Deng's work [9], they assume that: If the generic database is large enough, intuitively, one will find a generic subject p having similar face shape with gallery subject i , i.e., $R_p \approx R_i$, and then the illumination variation of subject i can be sparsely represented by that from generic subjects (illumination cone theory [10]). However, for a particularly face recognition task, designing such a generic face database is not easy. Moreover, the larger the generic database is, the larger the illumination variation dictionary is and the more computation cost (i.e., l^1 -minimization solver) we have. Without enough number of generic faces, the identity information contained in the reflectance part is difficult to be covered. We can conclude that under harsh illumination conditions where the illumination variation between test and gallery sample is large, with a relatively smaller size generic face database, the illumination variation dictionary constructed by pair-wise difference images cannot sparsely represent the illumination conditions, and have to resort to other gallery subjects, resulting in poor performance.

An intuitive illustration is given in Fig. 1, in which only one generic subject with 64 samples under varying illumination conditions are included. In this case, ESRC misclassified a test sample of subject 1 in harsh illumination condition to subject 14. We can also see that due to the existence of the identity information, several coefficients corresponding to the gallery dictionary have larger value than subject 1.

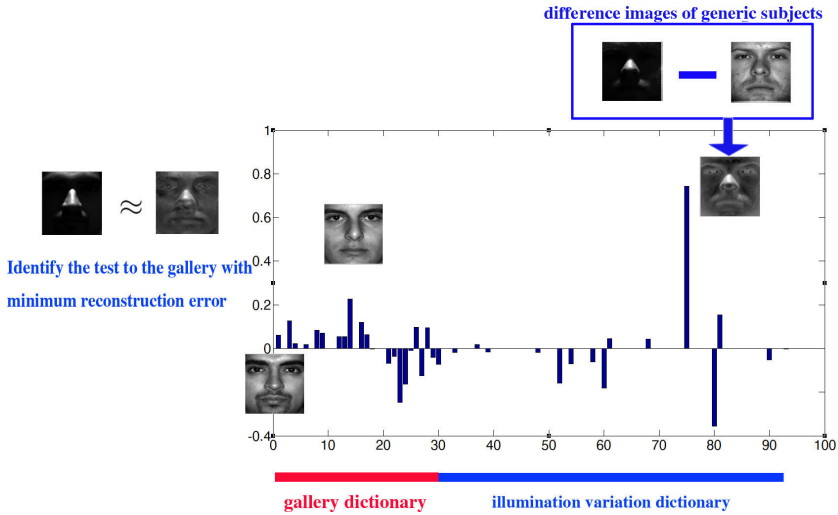


Fig. 1. Sparse representation for a sample face image with the help of illumination variation dictionary calculated from one generic subject with the samples under several illumination conditions. In this case, the test sample is misclassified because illumination variation dictionary calculated from generic database with small size is difficult to cover the identity variation.

3.2 Proposed Dictionary Designing

Based on the above analysis, to release the above assumption, an ideal illumination variation dictionary should exclude the identity information and retain the illumination variations only. This could be achieved from the following interesting observation:

With the assumption that the only variation between different samples from the same subject is caused by illumination, the identity information of a specific subject will be canceled out by the pair-wise difference images in logarithmic domain.

Proof

$$\begin{aligned}
 & \log F_{i,t} - \log F_{i,g} \\
 &= \log(R_i I_t) - \log(R_i I_g), \\
 &= \log(R_i) - \log(R_i) + \log(I_t) - \log(I_g), \\
 &= \log(I_t) - \log(I_g).
 \end{aligned} \tag{9}$$

Likewise we have $\log F_{p,1} - \log F_{p,0} = \log(I_1) - \log(I_0)$. Therefore, we can see that the difference images in logarithmic domain for face images of the same subject under the variant illumination conditions cancel out the identity information and retain the illumination variations only. This allows us to release the assumption of the existence of similar face shape in generic database in [9], and we can extend ESRC from “borrowing illumination variations from similar generic subjects”

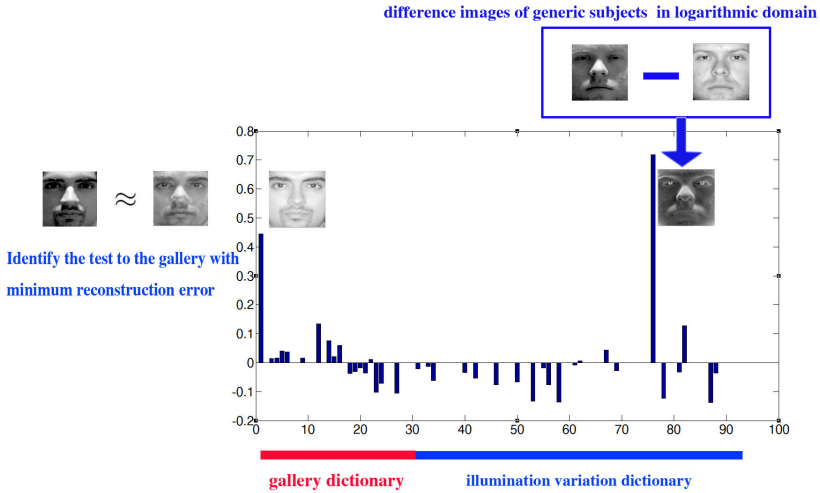


Fig. 2. Sparse representation for a sample face image with the help of illumination variation dictionary calculated *in logarithmic domain* from one generic subject with the samples under several illumination conditions. In this case, the test sample is correctly classified.

to “borrowing illumination variations from *any generic subject*”. Theoretically, with only one generic subject p , by capturing multiple samples under all possible illumination conditions and constructing the illumination variation dictionary: $A_e = [\log v_{p,1} - \log v_{p,0}, \log v_{p,2} - \log v_{p,0}, \dots, \log v_{p,N_p-1} - \log v_{p,0}]$, where N_p is the sample number for generic subject p . We can cover the illumination variations for all subjects and largely enhance the recognition performance. This interesting conclusion will be verified experimentally in the following section.

For the ease of discrimination, we will denote Deng’s original ESRC by ESRC(diff), and denote our proposed approach by ESRC(log-diff). From the aforementioned analysis, our proposed ESRC(log-diff) scheme requires generic face database with smaller size, and thus will be more computationally efficient than ESRC(diff). Moreover, the collection of the generic face database appears much easier. An intuitive illustration is given in Fig. 2, in which the same test sample as in Fig. 1 has been correctly recognized in this case. Note that the coefficients corresponding to the subject 1 clearly have much larger value compared with the other coefficients in gallery dictionary.

4 Experimental Results

4.1 Experiment Setting

In this section, experiments are conducted on two publicly available face databases with large illumination variations, namely, Extended Yale Face Database B [10] and FERET face database [15] to illustrate the effectiveness

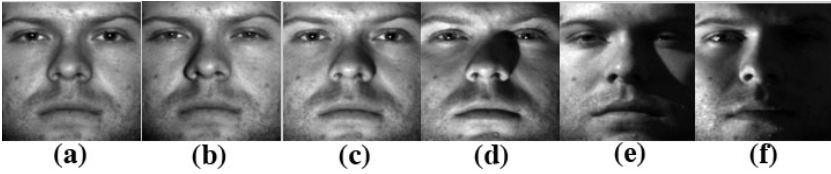


Fig. 3. Sample face images in Extended Yale B face database. (a) Gallery (b) Subset 1 (c) Subset 2 (d) Subset 3 (e) Subset 4 (f) Subset 5.

of our proposed method. All face images from the two databases are properly aligned, cropped and resized to 128×128 . To solve the l^1 -minimization problem, we adopt the Homotopy method (the matlab implementation are downloaded from [14]) with the error tolerance $\varepsilon = 0.05$.

We will also compare our method with three commonly used illumination preprocessing methods: Logarithmic transform (LOG) [13], Relative Gradient (RG) [12], WeberFace [11]. For these preprocessing based method, we use the nearest neighborhood rule with l^2 norm as the classifier. We also give the result of original images without any preprocessing (ORI) as the baseline.

4.2 Results on Extended Yale B

Extended Yale B face database includes 38 subjects under 9 poses and 64 illumination conditions. Only the frontal images were chosen in our experiments. Totally there are 2,414 frontal images of 38 subjects under 64 illumination conditions. They are divided into five subsets according to the angle between the light source directions and the central camera axis : subset 1 (0° to 12° , 263 images), subset 2 (13° to 25° , 456 images), subset 3 (26° to 50° , 455 images), subset 4 (51° to 77° , 526 images), subset 5 (above 78° , 714 images). Figure 3 shows samples of the same person from the five sets.

In our experiments, we partition the 38 subjects into two non-overlapped parts, the first 8 subjects, each with 63 difference images in pixel domain (ESRC(diff)) or logarithmic domain (ESRC(log-diff)) are the candidates for the illumination variation dictionary. For the remaining 30 subjects, images with the most neutral light condition ('A+00E+00') were used as the gallery, and the images from subset 3 – 5 were taken as the probes, since almost all methods achieve high recognition performance on subset 1 and 2. Figure 4 shows a plot of recognition rate versus the number of generic subjects on subset 3 – 5. As can be seen, when illumination condition is not so harsh, both ESRC(log-diff) and ESRC(diff) achieves high accuracy with a small number of generic subjects. When face images under harsh illumination conditions presented as probe (as in subset 4 and 5), ESRC(log-diff) largely outperform ESRC(diff). Particularly, compared with ESRC(diff), with only one generic subject, the recognition rates of ESRC(log-diff) for subset 4 and 5 notably increase from 49.28%, 8.54% to 84.06%, 71.89%, respectively.

The corresponding recognition rates of each methods for the three subsets are illustrated in Table 1. The recognition rates for ESRC(diff) and ESRC(log-

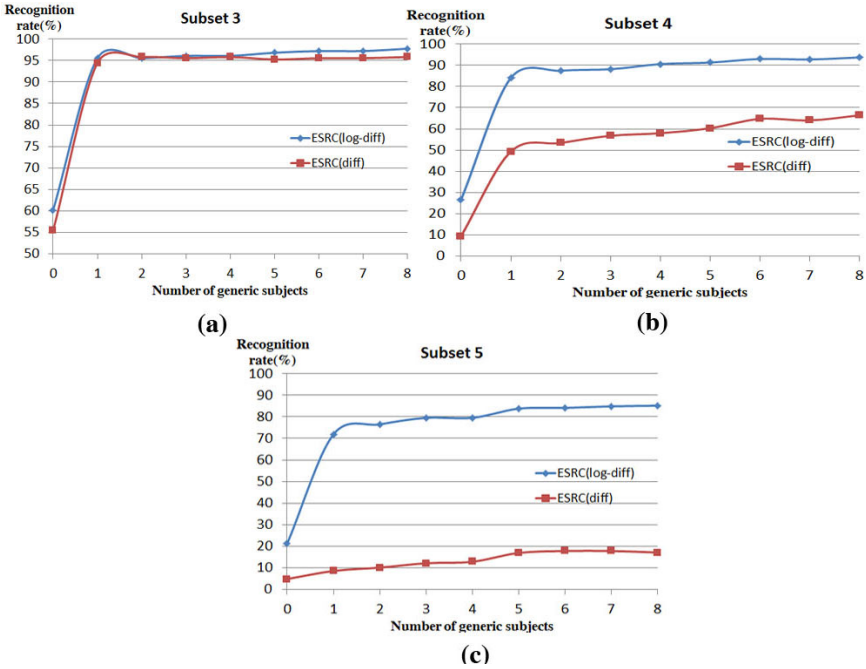


Fig. 4. The recognition rates versus the number of generic subjects for ESRC(log-diff) and ESRC(diff) on Subset 3-5 of Extended Yale B

diff) are given when all 8 generic subjects are used to construct the illumination variation dictionary. From the table, we could see that ESRC(log-diff) achieves the best performance, significantly better than SRC and ESRC(diff), even better than several representative preprocessing based methods.

Table 1. Recognition rates (%) on Extended Yale B

Methods	Subset 3	Subset 4	Subset 5	Average
$l_2(\text{ORI})$	49.03	10.87	4.98	18.65
$l_2(\text{LOG})$ [13]	54.87	24.40	24.02	32.43
$l_2(\text{RG})$ [12]	88.02	50.24	42.35	57.08
$l_2(\text{WeberFace})$ [11]	72.70	91.06	80.96	81.88
SRC [5]	55.43	9.42	4.30	19.85
ESRC(diff) [9]	95.82	66.43	17.08	53.56
ESRC(log-diff)	97.77	93.72	85.19	91.21

4.3 Results on FERET

The face images in Extended Yale B face database are captured in highly constrained conditions. To further testify our proposed method on practical applications, we conduct experiment on FERET database, one of the most commonly

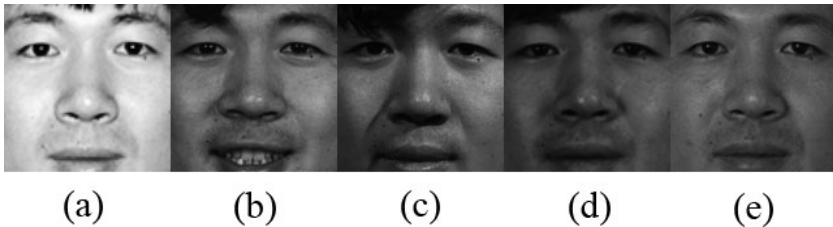


Fig. 5. Sample face images in FERET face database. (a) Fa (b) Fb (c) Fc (d) Dup I (e) Dup II.

used large-scale face database. We use the standard FERET protocol to conduct our experiments. The training set consists of 1,002 images of 429 subjects and is used as the generic database to construct the illumination variation dictionary. The gallery set Fa consists of 1,196 images of 1,196 subjects. There are four probe sets: Fb (different expressions with gallery, 1,195 images of 1,196 subjects), Fc (different illumination conditions with gallery, 194 images of 194 subjects), Dup I (images taken later in time, 722 images of 243 subjects), Dup II (images taken at least 18 months after the corresponding gallery, 234 images of 75 subjects). Figure 5 shows samples of the same person from the five sets.

We notice that the face images from Fc, Dup I, and Dup II have different illumination conditions with the gallery, and thus should benefit from our method. Table 2 illustrates the performance of SRC, ESRC(diff) and ESRC(log-diff), and the best results for FERET97 evaluation are given as a baseline. From the table we could see that our proposed method still performs the best.

Table 2. Performance evaluation of ESRC(log-diff) on FERET database

Methods	Fb	Fc	Dup I	Dup II
FERET97 Best [15]	96.0	82.0	59.0	52.0
SRC [9]	85.3	76.3	63.7	55.6
ESRC(diff) [9]	92.8	79.4	77.0	66.2
ESRC(log-diff)	92.8	83.0	78.5	70.9

5 Conclusion

The proposed illumination variation dictionary designing method is inspired from a simple observation from the Lambertian reflectance model: by calculating the difference image from two face images of the same subject under different illumination conditions in logarithmic domain, the identity information can be canceled out and only the illumination variation retains, which implies the illumination variation could be “borrowed” from *any generic subject*. To facilitate single-sample face recognition under varying illumination conditions, the illumination variation dictionary constructed in logarithmic domain is incorporated into the framework of ESRC. Face recognition experiments on Extended Yale B and FERET face databases demonstrate the superiority of our proposed method.

Acknowledgment. This work was supported by the Shenzhen-Hongkong Innovation Circle Project under grant No. ZYB200907070030A.

References

1. Zhao, W., Chellappa, R., Rosenfeld, A., Phillips, P.J.: Face Recognition: A Literature Survey. *ACM Computing Surveys*, 399–458 (2003)
2. Li, S.Z., Jain, A.K.: *Handbook of Face Recognition*, 2nd edn. Springer (2011)
3. Adini, Y., Moses, Y., Ullman, S.: Face Recognition: the Problem of Compensating for Changes in Illumination Direction. *IEEE Trans. Pattern Anal. Mach. Intell.* 19(7), 721–732 (1997)
4. Tan, X., Chen, S., Zhou, Z.H., Zhang, F.: Face Recognition from A Single Image per Person: A Survey. *Pattern Recognition* 39(9), 1746–1762 (2006)
5. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust Face Recognition via Sparse Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 31(2), 210–227 (2009)
6. Wagner, A., Wright, J., Ganesh, A., Zhou, Z., Mobahi, H., Ma, Y.: Towards A Practical Face Recognition system: Robust Alignment and Illumination by Sparse Representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 34(2), 372–386 (2012)
7. Chang, X., Zheng, Z., Duan, X., Xie, C.: Sparse Representation-Based Face Recognition for One Training Image per Person. In: Huang, D.-S., Zhao, Z., Bevilacqua, V., Figueroa, J.C. (eds.) *ICIC 2010*. LNCS, vol. 6215, pp. 407–414. Springer, Heidelberg (2010)
8. Chang, X., Zheng, Z., Duan, X., Xie, C.: Sparse Representation-Based Face Recognition for One Training Image per Person. In: Huang, D.-S., Zhao, Z., Bevilacqua, V., Figueroa, J.C. (eds.) *ICIC 2010*. LNCS, vol. 6215, pp. 407–414. Springer, Heidelberg (2010)
9. Deng, W.H., Hu, J., Guo, J.: Extended SRC: Undersampled Face Recognition via Intra-Class Variant Dictionary. *IEEE Trans. Pattern Anal. Mach. Intell.* (2012)
10. Georghiades, A., Belhumeur, P., Kriegman, D.: From Few to Many: Illumination Cone Models for Face Recognition Under Variable Lighting and Pose. *IEEE Trans. Pattern Anal. Mach. Intell.* 23(6), 643–660 (2012)
11. Wang, B., Li, W.F., Yang, W.M., Liao, Q.M.: Illumination Normalization Based on Weber’s Law With Application to Face Recognition. *IEEE Signal Process. Lett.* 18(8), 462–465 (2011)
12. Hou, Z., Yau, W.: Relative gradients for image lighting correction. In: *ICASSP 2010*, pp. 407–414 (2010)
13. Savvides, M., Kumar, V.: Illumination Normalization using Logarithm Transforms for Face Authentication. In: Kittler, J., Nixon, M.S. (eds.) *AVBPA 2003*. LNCS, vol. 2688, pp. 549–556. Springer, Heidelberg (2003)
14. <http://www.eecs.berkeley.edu/~yang/software/11benchmark>
15. Phillips, P.J., Moon, H., Rizvi, P., Rauss, P.: The Feret Evaluation Method for Face Recognition Algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(10), 1090–1104 (2000)

Efficient Extraction of Feature Signatures Using Multi-GPU Architecture

Martin Kruliš, Jakub Lokoč, and Tomáš Skopal

SIRET research group, Dept. of Software Engineering,
Faculty of Mathematics and Physics, Charles University in Prague
{krulis, lokoc, skopal}@ksi.mff.cuni.cz

Abstract. Recent popular applications like online video analysis or image exploration techniques utilizing content-based retrieval create a serious demand for fast and scalable feature extraction implementations. One of the promising content-based retrieval models is based on the feature signatures and the signature quadratic form distance. Although the model proved its competitiveness in terms of the effectiveness, the slow feature extraction comprising costly k-means clustering limits the model only for preprocessing steps. In this paper, we present a highly efficient multi-GPU implementation of the feature extraction process, reaching more than two orders of magnitude speedup with respect to classical CPU platform and the peak throughput that exceeds 8 thousand signatures per second. Such an implementation allows to extract requested batches of frames or images online without annoying delays. Moreover, besides online extraction tasks, our GPU implementation can be used also in a traditional preprocessing and training phase. For example, fast extraction allows indexing of huge databases or inspecting significantly larger parameter space when searching for an optimal similarity model configuration that is optimal according to both efficiency and effectiveness.

Keywords: similarity search, feature extraction, GPU, parallel.

1 Introduction

The traditional approaches to the multimedia retrieval rely on the well-established fulltext search. However, as the amount of new multimedia data grows immensely nowadays, there often appear scenarios where data annotations cannot be provided and so the content-based retrieval techniques in connection with the similarity search paradigm is the only viable possibility for computer-aided image retrieval [8]. The content-based retrieval approach requires an effective similarity model that should mimic a user's perception of which images are similar and which are not [22]. More specifically, the similarity model consist of features extracted from the original images (formed into image descriptors) and a total similarity function (often modeled as a distance function) providing a similarity ranking (ordering) on the descriptors. The similarity model is effective if the ranking corresponds to user preferences, that are often provided in the form of so-called ground truth testbed.

For example, the ground truth can be represented as an annotated sample of the database, and is often utilized to train or verify the similarity model [7].

While the traditional content-based retrieval applications use a slow preprocessing phase to prepare feature representations, recent applications calls for faster feature extraction tools enabling immediate online analysis of the picture content. Examples of such systems can be, e.g., camera systems producing a lot of video streams that have to be processed frame by frame or image exploration techniques that create exploration structures from a given set of images during the interaction with users [14]. Furthermore, the fast feature extraction can be utilized also in traditional tasks, e.g., for fast indexing of huge datasets or for training of similarity models, where significantly larger parameter space can be considered and tested.

In this paper, we focus on the similarity models based on feature signatures, that can be effectively compared via the *signature quadratic form distance* (SQFD) [4,6]. This model is competitive in the terms of the effectiveness and at the same time provides many parameters to tune. We provide efficient GPU implementation of the expensive feature extraction process resulting in two orders of magnitude speed up. The contributions of this paper can be summarized as:

- Description of the feature signatures extraction process with emphasis on performance issues.
- Proposal of a fast GPU implementation of this extraction process.
- Experimental results expressing the power of our GPU feature extraction.

The paper is organized as follows. In Section 2, we describe the similarity model based on feature signatures and the basics of feature extraction process. In Section 3, we recall GPU platform basics and describe details of our GPU implementation of the feature extraction process. The experimental results are presented and discussed in Section 4 and Section 5 concludes the paper.

2 Similarity Model Based on Feature Signatures

In this section, we sketch the recently studied similarity model based on feature signatures and the Signature Quadratic Form Distance, that can be utilized to solve the content-based image retrieval tasks [5,3]. Especially, we describe in detail the employed feature extraction method, where we highlight parameters influencing many aspects of the resulting similarity space. We also remember recent works focusing on efficient query processing in this model – the metric and ptolemaic indexing of the Signature Quadratic Form Distance.

2.1 Feature Signatures

The traditional object representation approaches utilizing feature histograms aggregate features within predefined bins of fixed-sized vectors. Unlike the feature histograms, the feature signatures allow a more flexible object representation

within a utilized feature space [9,18,8], where the size of resulting feature signatures is not fixed. Hence, complex multimedia objects can be represented by a feature signature consisting of many centroids, while simple multimedia objects have just few centroids in their feature signatures.

Definition 1 (Feature Signature). *Given a feature space \mathbb{F} , the feature signature S^o of a multimedia object o is defined as a set of tuples from $\mathbb{F} \times \mathbb{R}^+$ consisting of representatives $r^o \in \mathbb{F}$ and weights $w^o \in \mathbb{R}^+$*

In order to compare feature signatures, the SQFD, which is a generalization of the conventional QFD, is employed. In contrast to the well-known Earth Mover’s Distance, the SQFD makes it possible to balance the tradeoff between indexability and retrieval quality [2]. The authors have demonstrated that the parameters of the similarity functions affect the indexability of the underlying data space, thus allowing to balance the tradeoff between indexability and retrieval quality. It was shown that even a very simple metric pivot table approach [23] can reach a speedup factor of up to 170 with respect to the sequential scan. In addition, the combination of the SQFD and ptolemaic pivot tables has shown a speedup factor of up to 300 [15]. In the meantime, Krulis et al. [13] came up with the idea of processing the SQFD on many-core GPU architectures. By implementing the query evaluation process on many-core GPUs and also multi-core CPUs, they have shown a significant improvement in efficiency compared to the serial approaches.

2.2 Extraction of the Feature Signatures

The feature extraction process determining a feature signature from an image consists of several consecutive steps, each of them providing several options with various sets of parameters. The basic overall schema for this process is depicted in Figure 1. The image is preprocessed first by common image algorithms. After that, a suitable sampling method is selected and several features are extracted for every sampled point. Finally, all features are clustered via the k-means clustering. In the following paragraphs, we explain details of the feature extraction and the k-means algorithm used in our implementation.

In the feature extraction step, the sampled points are mapped into the requested feature space \mathbb{F} . We utilize seven-dimensional representatives $f_i^o = (x, y, L, a, b, c, e) \in \mathbb{F} \subseteq \mathbb{R}^7$, where (x, y) are the coordinates of the sampled point, (L, a, b) represent the color of the sampled point mapped into the CIE Lab color space [12], and (c, e) are contrast and entropy values computed from the neighborhood of the point in the corresponding gray-scale image. We compute texture information from the gray level co-occurrence matrix G [10] extracted from the neighborhood of the point, where each point is assigned an intensity $i \in I^1$. Since the value ranges of utilized features differ significantly, we also normalize values from each dimension into $[0, 1]$ interval.

¹ More specifically, the contrast c and entropy e are evaluated as $c = \sum_{i,j \in I} (i - j)^2 \times G(i, j)/n$ and $e = - \sum_{i,j \in I} (G(i, j)/n) \times \log(G(i, j)/n)$, where $n = \sum_{i,j \in I} G(i, j)$.

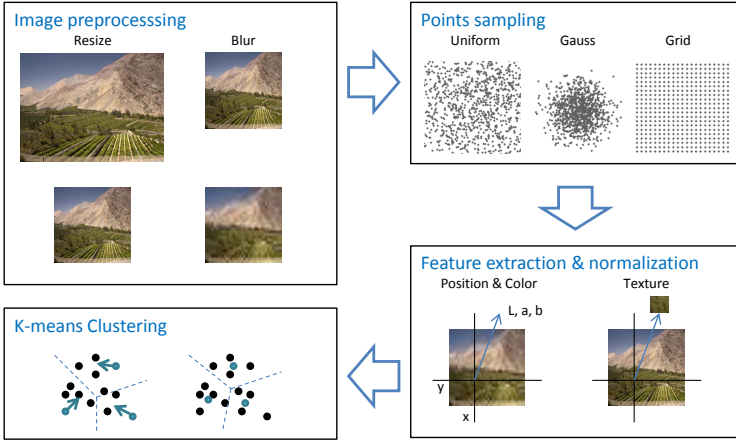


Fig. 1. Extraction schema for feature signatures

In the last step of the overall feature extraction process, all the extracted representatives $f_i^o \in \mathbb{F}$ are aggregated using the k -means clustering algorithm [17] employing the weighted L_p norm distance. The weights change the impact of each utilized feature and thus fundamentally influence the result of the clustering. By setting a weight to zero, we can even totally ignore the effect of a particular feature. The k -means clustering is an iterative method, where the number of iterations is defined by the user. In each iteration, all representatives $f_i^o \in \mathbb{F}$ are distributed within the actual set of centroids $r_i^o \in \mathbb{F}$ of the clusters $\mathcal{C}_i^o \subseteq \mathbb{F}$. Then, for each cluster a new centroid is created by averaging all the points in the cluster, which can be depicted as a shift of the old centroid to the clusters center of gravity. When using an adaptive version of the k -means clustering, we can also remove some too close or too small clusters and thus influence the number of the resulting clusters. As a result of the k -means clustering algorithm, the feature signature consisting of representatives $r_i^o \in \mathbb{F}$ and weights $w_i^o \in \mathbb{R}^+$ is created, where each representative $r_i^o \in \mathbb{F}$ corresponds to the centroid of the cluster $\mathcal{C}_i^o \subseteq \mathbb{F}$ obtained in the last iteration of the k -means, i.e.,

$$r_i^o = \frac{\sum_{f \in \mathcal{C}_i^o} f}{|\mathcal{C}_i^o|}, \text{ with relative frequency } w_i^o = \frac{|\mathcal{C}_i^o|}{\sum_i |\mathcal{C}_i^o|}.$$

3 Implementation

Before we introduce the details of our GPU extractor, let us briefly revise current GPU architecture. Our implementation was developed and tested on the NVIDIA Fermi architecture [20], however, it should work on the new Kepler architecture as well as on the current AMD GPU devices. GPU architectures differ from CPU architectures in multiple ways. The most important two are rather specific thread execution model and complex memory model. The CPU is designed so that each core process one independent thread at a time. Threads

running on GPU all execute the same program and small groups of threads even execute the same instruction at a time (Single Instruction Multiple Threads).

The GPU is a rather independent device, so it has its own memory. This means that all input data must be transferred to the device and computed results must be transferred back to memory of the host system. Furthermore, there are multiple types of memories – *the global memory* (of several GBs), *the local (or shared) memory* (tens of kBs), and the *private memory* (registers of each core). Each memory has some specific limitations which are inevitably inherited from the parallel nature of the architecture.

We would like to summarize some of the architecture implications and best programming practices suggested by the vendor [19]:

- The latency of data transfers between the host system and the GPU devices needs to be inhibited. Therefore, we should bulk the transfers and try to overlap them with GPU computations.
- Data structures must be designed according to memory limitations of the GPU. The data placement must be considered carefully as different types of memories have different properties (especially the size and speed).
- The algorithm must embrace the SIMT execution model, at least for the parts of the work being processed by one thread group. This usually requires significant modifications of the algorithm or selection of a different algorithm solving the same problem.
- A multitude of threads (at least thousands) needs to be spawned in order to utilize all available cores and balance the load efficiently.

3.1 GPU Extractor

Our GPU extractor implementation is quite complex. In this section, we will focus on the key details that allowed us to achieve such excellent performance.

The extractor exploits two approaches to parallelism. Each signature is computed by a SIMT parallel algorithm and multiple signatures are computed concurrently. The CPU code ensures the loading of images from persistent data store, create blocks of images of appropriate size, and dispatch these blocks to the available GPUs. Each block is transferred to the GPU, then the GPU computes signatures for all images in the block, and the block of signatures is transferred back to the host memory. The blocks are dispatched so there are always two blocks assigned to one GPU. One block is being computed while the data of the other block are transferred. Furthermore, there are two CPU threads allocated for each GPU device. These threads are responsible for feeding the GPU, waiting for the GPU to terminate, and consolidation of the results.

Images in the block are processed as depicted in Figure 2. Each thread group (assigned to one SMP²) processes one image and all threads in the group synergically cooperates to compute the signature. The optimal block size was empirically determined as $2 \times$ number of SMPs on the GPU (in our case $2 \times 16 = 32$).

² Symmetric Multi-Processor unit, which contains 32 synchronously running cores.

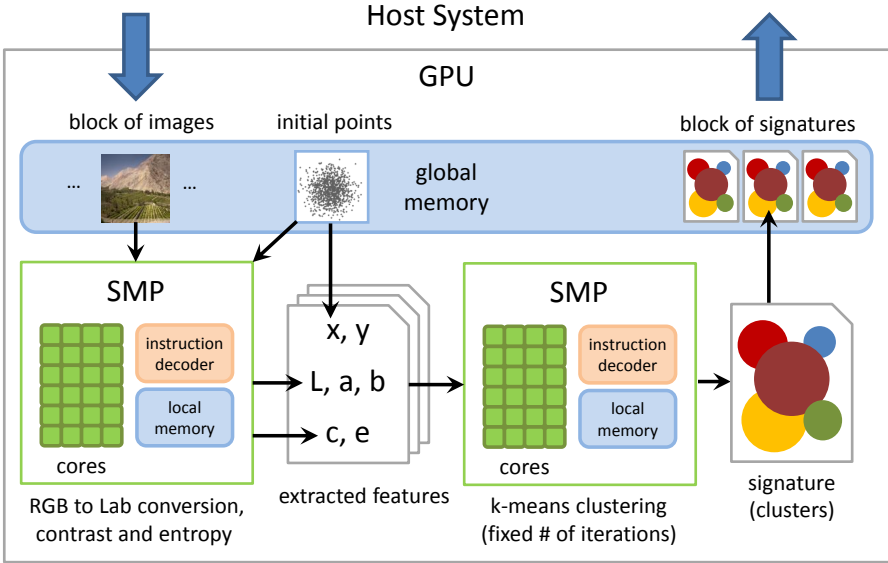


Fig. 2. Schema of the extraction process of one image

Different approaches are clearly suboptimal. Computing multiple signatures by one thread group would be highly complicated as the size of the local memory is very limited and its utilization is very important to overall performance of both feature extraction and k-means clustering. Computing one signature by multiple groups would be even more complicated as the groups have limited means of communication and synchronization.

Feature Extraction Process. The first phase of the signature creation is the feature extraction and normalization process (see Section 2.2). The sampling points (whole set) are provided by CPU and uploaded into constant global memory before the extraction is started. The extraction of the first 5 dimensions of the feature space (x, y, L, a, b) is quite straightforward. The (x, y) coordinates are computed from the sampling points coordinates as a simple linear combination. The RGB value of the corresponding pixel is taken and converted into the CIE Lab color space by transformation equations from the CIE Lab specification.

The computation of contrast and entropy features is slightly more complicated. The bitmap is converted to gray-scale using all threads in the group. Since each pixel is represented with only a few bits and the GPU natively processes data in 32-bit words, we use simple bit-packing technique that stores multiple pixels in one word. We convert the entire bitmap and keep it in the local memory of the SMP for the sake of simplicity and parallelism even though only sampled pixels and their surroundings are required for the computation.

To compute contrast and entropy, the co-occurrence matrix G must be constructed for each pixel. The matrix is rectangular $|I| \times |I|$, where I is the set of

possible intensities. Since we use 4-bit gray-scale in our experiments, the $|I| = 16$ and the matrix has 16×16 items. We allocate as many matrices as possible in the local memory and assign one thread to each matrix. These threads iteratively process initial points, construct corresponding co-occurrence matrices, and compute contrast and entropy. The time complexity of this step depends on the size of the matrix G and size of the neighborhood of the pixel, which are constant for all points. Therefore, each thread performs almost the same amount of work.

Described algorithm was designed under the assumption that it is possible to fit at least as many G matrices to local memory (along with the gs-bitmap), as there are cores on SMP, or (better) as there are threads in the corresponding group. This assumption holds in our case³ as we are able to accommodate approximately 150 matrices in the local memory. On the other hand, if we use more bits per pixel in the gray-scale bitmap, the bitmap itself and the matrices will be significantly larger and a different algorithm could be more suitable for the problem.

K-means Clustering. The second phase is the k-means clustering performed on the points from a feature space (see Section 2.2). Since the k-means algorithm has many variations, we need to specify several details:

- We use fixed number of iterations and this number is a configurable parameter of the extractor. This way more complex images end up with more centroids in their signatures than the simple images.
- The clusters that have centroids closer than specified threshold (which is also a parameter) are merged together.
- After each iteration, clusters that are smaller than $s \times i$, where i is the number of the iteration and s is a parameter of the extractor, are thrown away. Points from these clusters are not dismissed, but rather reassigned in next iteration.
- Our algorithm does not care for the final assignment of points to clusters, but only for the final centroids and weights (number of points in each cluster).

All the threads in the group follow the algorithm steps together waiting on an explicit barrier after each step of the algorithm. One k-means iteration consists of the following steps:

1. The closest centroid is found for each point and coordinates of the point are atomically added to the new centroid coordinates (per dimension). Also the weight of the closest centroid is atomically incremented.
2. New centroid coordinates are computed dividing sums from previous step by number of points in the corresponding cluster (computing an average).
3. The clusters with centroids closer than joining threshold are merged.
4. The clusters smaller than $s \times i$ limit are disposed of.

First two steps are embarrassingly parallel. Each point may be processed independently and we assume that there are more points than threads in a group.

³ SMP has 48 kB of local memory, thumbnails are 150×150 px in 4-bit gray-scale.

The only interesting issue is the optimal data representation. We represent each set of N d -dimensional points⁴ as d arrays of N values rather than an array of N structures with d values. This representation better fits the properties of both global memory and local memory where the points and centroids are stored.

It is possible to use kd-trees or other geometric data structures to accelerate the nearest neighbour problem (the first step) [11]. We can also use an approximate approach to k-means [21]. However, these techniques are faster only for large number of centroids, and kd-trees do not perform well in the GPU memory. As we use only hundreds of centroids, empirical results show that it is better to pursue raw power of parallelism with the simplest algorithm.

The third step tests the distance of every centroid pair. The centroid pairs are iterated using the nested two-level for-loop, where only the inner loop is parallelized and the explicit barrier synchronization is performed after each iteration of the outer loop. This way the parallelism is slightly reduced, however, we do not require any means of data synchronization. In order to avoid expensive merging and array compacting, this step just sets the weight of one of the merged clusters to 0, so the cluster will be disposed of by the last step. More elaborate methods did not show any measurable speedup as the third step is significantly cheaper than the remaining steps.

The last step filters out small clusters and compacts the set of centroids, so the arrays does not contain empty elements. First, the compacting step computes new offset for each nonempty element. The offsets are computed by standard binary reduction tree algorithm performed by all available threads. Finally, all nonempty elements are copied into new compacted array at their new positions.

4 Experimental Results

In the experiments we focused on the efficiency of the extraction and the MAP evaluation process using GPU architecture, where we measured the speed up according to the CPU platform.

4.1 The Testbed

The *Thematic Web Images Collection* (TWIC) database of 11,555 images divided in 200 classes [16] was employed for basic experiments and approximately 17.5 mil. images from the profimedia dataset [1] were used for large-data tests.

Our experiments were performed on a server with special motherboard (FT72-B7015) designed to embrace up to 8 GPUs. The server was equipped with Xeon E5645 processor that contains 6 physical (12 logical) cores running at 2.4 GHz, 96 GB of DDR3-1333 RAM, and 4 NVIDIA Tesla M2090 GPU cards (Fermi architecture). Each GPU chip has 512 cores (32 cores per 16 SMPs) and 6 GB of memory. We also tested the implementation on commodity PC with two NVIDIA GTX 580 which have also 512 cores, but only 1.5 GB of memory. We have found that in our case the gaming GTX 580 cards have similar performance as the much more expensive Tesla cards, thus we do not provide detailed comparison.

⁴ As described in 2.2, the $d = 7$ in our case.

4.2 Performance Tests

Since the major contribution of this paper is a fast GPU extraction implementation, we provide performance test results and comparison to the CPU extractor. All the times were measured using the system real-time clock. We are aware that this method is not entirely precise and there are many both technical and philosophical issues regarding performance benchmarks. However, these tests are designed to give the reader a general idea about the performance rather than provide an accurate comparison. All tests were conducted using parameters that produced the highest precision results.

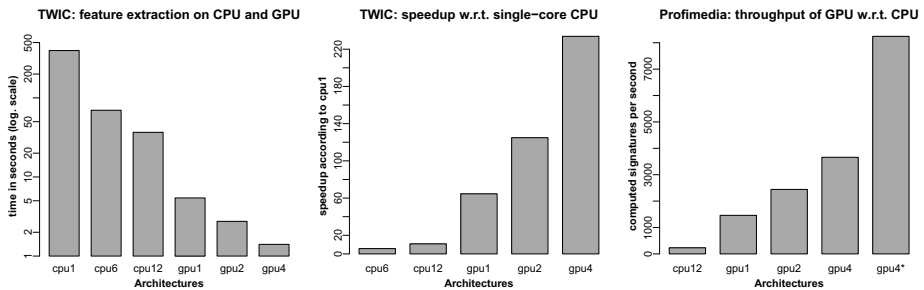


Fig. 3. Time, speedup, and throughput comparisons

Figure 3 summarizes the times and speedup of the experiments conducted on CPU (using different numbers of threads) and GPU (using different numbers of devices). The *cpu t* methods designate tests running on CPU with t threads⁵ and the *gpu d* methods designate tests running on d GPU devices. The times depicted in the first graph are separated into two columns – the extraction process of 11,555 images (TWIC dataset [16]), which we used for tests designed to explore the parameter space of the extractor.

The tests were evaluated $233\times$ faster on 4 GPUs than on single-core CPU and $21.7\times$ faster than on 12 core CPU. Thanks to this speedup, we were able to conduct all experiments presented in previous section in the matter of hours. The same experiments would take days on 12 CPU cores and weeks on single-core.

We have also considered using the GPU extractor for the indexing and video stream processing techniques. For testing purposes, we have extracted a database of 1,000,000 images [1] and created signature index by both GPU and CPU extractors. The extractor running on 4 GPUs can extract 8244 signatures per second while on 12 CPU cores the throughput is only 303 signatures per second. These tests were conducted so that all images were pre-cached in RAM, hence the extractor has not been slowed down by loading data from persistent storage. Since each image has size of approximately 33.9 KB, the system would require a persistent storage that is capable of reading data at the minimum rate of 273.3 MB/s, which cannot be easily achieved by common hard disk drives. Previous

⁵ The tests run up to 12 threads as we have 12 core CPU.

tests were focused on the speed of the extractor. If the extractor is employed as an indexing service, the performance of the persistent data storage cannot be ignored. We have equipped our server with two RAID 0 arrays both containing two common hard disks. One array kept the input images and the other array was used for storing signatures. We have used the same data source as for the previous experiment, but we took 17.5 millions of images to ensure that the data nor the result will fit the RAM. The throughput of the extractor is depicted in the third graph of the Figure 3. The speed of the extractor has dropped to 3661 signatures per second on four GPUs. It is $2.25\times$ slower than previous experiment, in which the images were cached in RAM (denoted `gpu4*` in the graph). Based on the empirical data, we speculate that the feature extraction system would require at least 4 modern SSD drives connected to RAID 0 in order to match the speed of 4 GPU devices.

5 Conclusions

In this paper, we present a highly efficient GPU implementation of the feature extraction of image signatures, reaching more than two orders of magnitude speedup with respect to classical CPU platform. It also achieves a throughput of 8244 signatures per second, which is far beyond the throughput of common hard drives or network devices. Fast feature extraction implementation is critical in many recent applications like video stream processing or image exploration techniques, where for user interaction scenarios the low response times are essential. Our GPU implementation can be used also in traditional indexing or training tasks, where huge datasets have to be extracted or broad parameter spaces have to be inspected.

Acknowledgments. This research has been supported in part by Czech Science Foundation projects P202/11/0968, P202/12/P297 and Charles University grant agency (GAUK) project 277911.

References

1. Profimedia Image Database, <http://www.profimedia.cz/>
2. Beecks, C., Lokoč, J., Seidl, T., Skopal, T.: Indexing the signature quadratic form distance for efficient content-based multimedia retrieval. In: Proc. ACM Int. Conf. on Multimedia Retrieval, pp. 24:1–24:8 (2011)
3. Beecks, C., Seidl, T.: Analyzing the inner workings of the signature quadratic form distance. In: Proc. IEEE Int. Conference on Multimedia and Expo. (2011)
4. Beecks, C., Uysal, M.S., Seidl, T.: Signature quadratic form distances for content-based similarity. In: Proc. 17th ACM Int. Conference on Multimedia (2009)
5. Beecks, C., Uysal, M.S., Seidl, T.: A comparative study of similarity measures for content-based multimedia retrieval. In: Proc. IEEE International Conference on Multimedia & Expo. (2010)
6. Beecks, C., Uysal, M.S., Seidl, T.: Signature quadratic form distance. In: Proc. ACM International Conference on Image and Video Retrieval, pp. 438–445 (2010)

7. Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large scale online learning of image similarity through ranking. *J. Mach. Learn. Res.* 11, 1109–1135 (2010)
8. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.* 40(2), 5:1–5:60 (2008)
9. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: an experimental comparison. *Information Retrieval* 11(2), 77–107 (2008)
10. Gotlieb, C.C., Kreyzig, H.E.: Texture descriptors based on co-occurrence matrices. *Comput. Vision Graph. Image Process.* 51(1), 70–86 (1990)
11. Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., Silverman, R., Wu, A.: An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(7), 881–892 (2002)
12. Kasson, J.M., Plouffe, W.: An analysis of selected computer interchange color spaces. *ACM Trans. Graph.* 11(4), 373–405 (1992)
13. Krulis, M., Lokoc, J., Beecks, C., Skopal, T., Seidl, T.: Processing the signature quadratic form distance on many-core gpu architectures. In: *CIKM*, pp. 2373–2376 (2011)
14. Lokoč, J., Grošup, T., Skopal, T.: Image exploration using online feature extraction and reranking. In: *Proceedings of the 2nd ACM Int. Conference on Multimedia Retrieval, ICMR 2012*, pp. 66:1–66:2. ACM, New York (2012)
15. Lokoč, J., Hetland, M.L., Skopal, T., Beecks, C.: Ptolemaic indexing of the signature quadratic form distance. In: *Proceedings of the Fourth International Conference on Similarity Search and Applications (SISAP)*. Springer (2011)
16. Lokoč, J., Novák, D., Skopal, T., Sibirkina, N.: Thematic Web Images Collection, SIRET Research Group (2012), <http://siret.ms.mff.cuni.cz/twic>
17. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability* (1967)
18. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(10), 1615–1630 (2005)
19. NVIDIA. CUDA Programming Guide, http://developer.download.nvidia.com/compute/DevZone/docs/html/C/doc/CUDA_C_Programming_Guide.pdf
20. NVIDIA. Fermi GPU Architecture, http://www.nvidia.com/object/fermi_architecture.html
21. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007*, pp. 1–8. IEEE (2007)
22. Tversky, A.: Features of similarity. *Psychological Review* 84(4), 327–352 (1977)
23. Zezula, P., Amato, G., Dohnal, V., Batko, M.: Similarity Search: The Metric Space Approach. In: *Advances in Database Systems*, Springer-Verlag New York, Inc., Secaucus (2005)

Collaborative Tracking: Dynamically Fusing Short-Term Trackers and Long-Term Detector

Guibo Zhu, Jinqiao Wang, Changsheng Li, and Hanqing Lu

National Laboratory of Pattern Recognition, Institute of Automation,
Chinese Academy of Sciences, Beijing 100190, China
{gbzhu, jqwang, csl, luhq}@nlpr.ia.ac.cn

Abstract. This paper addresses the problem of long-term tracking of unknown objects in a video stream given its location in the first frame and without any other information. It's very challenging because of the existence of several factors such as frame cuts, sudden appearance changes and long-lasting occlusions etc. We propose a novel collaborative tracking framework fusing short-term trackers and long-term object detector. The short-term trackers consist of a frame-to-frame tracker and a weakly supervised tracker which would be updated under the weakly supervised information and re-initialized by long-term detector while the trackers fail. Additionally, the short-term trackers would provide multiple instance samples on the object trajectory for training a long-term detector with the bag samples with P-N constraints. Comprehensive experiments and comparisons demonstrate that our approaches achieve better performance than the state-of-the-art methods.

Keywords: collaborative tracking, online learning, samples selection.

1 Introduction

Long-term tracking in unconstrained environments is a very active topic in computer vision due to its wide-ranging applications in video indexing, surveillance, human-computer interaction, augmented reality, etc. [1, 2]. A tracking system usually consists of three components: 1) an appearance model, used for evaluating the likelihood that the object of interest is at some particular location; 2) a motion model, which relates the locations of the object over time; 3) a search strategy for finding the most possible location in the current frame [3]. However, the problem and difficulty in a tracking system depend on several sources of varieties such as changes in appearance, varying lighting conditions, cluttered background, partial or complete occlusion, and frame-cuts.

Nowadays, various tracking algorithms have been proposed [13, 14, 17, 4, 6]. Template tracking [13, 14, 15] is the most straightforward approach that estimates the objects' motion between consecutive frames. Templates have limited modeling capability as they represent only a single appearance of the object. To deal with more appearance variations, the generative models [17, 18, 19, 20, 21] have been proposed. However, the generative trackers only model the appearance of the object and as such often fail in cluttered background. In order to alleviate this problem, training an

adaptive discriminative classifier in an online manner to distinguishing the object from the background has shown promising results [3, 4, 5, 6]. The essential phase of adaptive discriminative trackers is the update: the close neighborhood of the current location is used to sample positive training examples, distant surrounding of the current location is used to sample negative examples, and these are used to update the classifier in each frame. It has been demonstrated that this updating strategy handles significant appearance changes, short-term occlusions, and cluttered background. However, these methods suffer from drift and failure if the object leaves the scene for a long time. To address the problems, the update of the tracking classifier has been constrained by an auxiliary classifier trained in the first frame [7] or by training a pair of independent classifiers [8, 9].

In this paper, we focus on the problem of long-term tracking an arbitrary object with no prior knowledge other than its location in the first frame. To develop a robust updating adaptive appearance models, we would like to handle partial occlusions or disappearance without significant drift through exploring the interrelationship between the short-term tracker and the long-term detector. Here, the adaptive short-term trackers consist of a frame-to-frame tracker and a weakly supervised tracker which would be updated under the weakly supervised information and re-initialized by long-term detector while the trackers fail. Simultaneously, the adaptive short-term trackers would provide multiple instance samples on the object trajectory for training a long-term detector. Unlike previous methods, we exploit the steady local information of object and develop the adaptive short-term trackers. Our algorithm dynamically fuses adaptive trackers and detector, which can deal with the appearance model and the motion model in a novel framework. Experimental results on the public available datasets demonstrate the effectiveness of our method.

The rest of the paper is organized as follows. In the next section, we introduce our tracking algorithm; in Section 2, we present qualitative and quantitative results of our tracker on a number of challenging image sequences. We draw the conclusion in Section 4.

2 The Proposed Approach

We present details of the robust visual tracking framework by fusing adaptive short-term trackers and long-term detector, as shown in Fig. 1.

The components of the framework are characterized as follows: the frame-to-frame tracker estimates the object's motion between consecutive frames. Adaptive short-term tracker estimates the object's location under the assumption that the object is visible or partial visible. If the object moves quickly or is occluded partially abruptly, the adaptive short-term tracker may recover when the frame-to-frame tracker is likely to fail and never recover by itself. The adaptive short-term trackers could provide multiple instance samples on the object trajectory for training a long-term detector.

The trained detector will scan full of the frame to localize all possible candidate patch that is similar to all appearances observed. Learner evaluates the performance of trackers and detector, estimates detector's errors and generates the credible templates

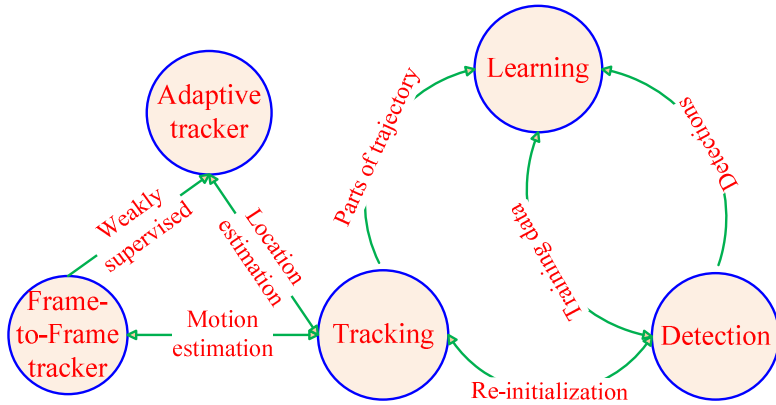


Fig. 1. The block diagram of our approach

and training data. The training data consists of bag samples to reinforce the detector's capability. For alleviating the effect from the condition that both the frame-to-frame tracker and the detector fail, we introduce adaptive short-term tracker and P-N constraints for bag samples selection to improve the detector's generalization capability. Additionally, because of the existence of object templates learned from the past, the learning strategy makes the detector have strong ability to discriminate the object against background.

2.1 Short-Term Trackers

Adaptive short-term trackers contain a frame-to-frame tracker and an approximate multiple instance learning tracker (MIL) [3]. Frame-to-frame tracker is used for exploring the motion of consecutive frames. We adopt the approach of Kalal et al. [21] for recursive tracking which bases on Lucas-Kanade tracker (KLT) [13].

The approach of KLT bases on three assumptions. The first assumption is referred to as brightness constancy [23] and is

$$I(X) = J(X + d) \quad (1)$$

Eq. (1) states that a pixel at the two-dimensional location X in an image I might change its location in the second image J but retains its brightness value. The vector d will be referred to as the displacement vector. The second assumption is referred to [22] as temporal persistence. It states that the displacement vector is small. Small in this case means that $J(X)$ can be approximated by

$$J(X) \approx I(X) + I'(X)d \quad (2)$$

where $I'(X)$ is the gradient of I at location X .

The third assumption, known as spatial coherence, alleviates this problem. It states that all the pixels within a window around a pixel move coherently. By incorporating this assumption, d is found by minimizing the term

$$\sum_{(x,y) \in W} (J(X) - I(X) - I'(X)d)^2 \quad (3)$$

which is the least-squares minimization of the stacked equations. The size of W defines the considered area around each pixel. Additional implementation details are in [22].

According to the forward-backward error measure [21], Lucas-Kanade method is applied twice on points B_{b1} in the bounding box of the object and measured based on the similarity of the patches R surrounding points B_{b1} and the patches P_2 surrounding the tracked points B_{b2} . Since the normalized correlation coefficient is invariant against uniform brightness variations [23], the similarity of these two patches R and P_2 is calculated by the Normalized Correlation Coefficient (NCC) as

$$NCC(P_1, P_2) = \frac{1}{n-1} \sum_{x=1}^n \frac{(P_1(x) - \mu_1)(P_2(x) - \mu_2)}{\delta_1 \delta_2} \quad (4)$$

where μ_1, μ_2, σ_1 and σ_2 are the means and standard deviations of R and P_2 .

Under the three assumptions of Lucas-Kanade, this frame-to-frame tracker could provide samples for the long-term detector. If any of the three assumptions are not met, the frame-to-frame tracker would have a failure so that it couldn't provide the enough training samples for long-term detector, which has enormous influence on the tracked results. If the object is occluded quickly, the assumptions will be violated. For solving this problem, we introduce a weakly supervised tracker which could mine the discriminative local patch information and estimate the object effectively in short term, especially when the long-term detector isn't trained sufficiently.

In this paper, we use weakly supervised multiple instance learning tracking (WSMILT) as our weakly supervised tracker. Unlike MIL Track [3], WSMILT will use the weakly supervised information from frame-to-frame tracker. The basic flow of adaptive short-term tracker in this work is illustrated in Fig.1 and summarized in Algorithm 1. Like MIL Track [3], we extract a set of Haar-like features for each image patch [11, 24]. Then the appearance model is composed of a discriminative classifier which is able to return $p(y = 1 | x)$, where x is an image patch and y is a binary variable indicating the presence of the object of interest in that image patch. At every time step t , our weakly supervised tracker maintains the object location l_t^* . Let $l(x)$ denote the center location of image patch x . For each new frame, if the frame-to-frame tracker has tracked the object, we crop out a set of image patches $X_s = \{x : \|l(x) - l_{t-1}^*\| < s\}$ that are within some search radius s of the current tracker location, and compute $p(y = 1 | x)$ for all $x \in X^s$. We then use a greedy strategy to update the tracker location:

$$l_t^* = l \left(\arg \max_{x \in X^s} p(y=1|x) \right) \quad (5)$$

In other words, we don't maintain a distribution of the target's location at each frame, and our motion model assumes that the location of the tracker at time t is equally likely to appear within a radius s of the tracker location at time $(t-1)$:

$$p(l_t^* | l_{t-1}^*) \propto \begin{cases} 1 & \text{if } \|l_t^* - l_{t-1}^*\| < s \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

Algorithm 1. Weakly Supervised Multiple Instance Learning Tracking

Input: Video frame number k

Method:

1: Crop out a set of image patches, $X_s = \{x : \|l(x) - l_{t-1}^*\| < s\}$ and compute feature vectors.

2: Use multiple instance learning classifiers to estimate the probability $p(y=1|x)$ for $x \in X^s$.

3: Update the tracker location $l_t^* = l \left(\arg \max_{x \in X^s} p(y=1|x) \right)$.

4: Crop out two sets of image patches $X^r = \{x : \|l(x) - l_t^*\| < r\}$ and $X^{r,\beta} = \{x : r < \|l(x) - l_t^*\| < \beta\}$, where $r < s < \beta$.

5: If the frame-to-frame tracker has tracked the object, we update MIL appearance model with one positive bag X^r and $|X^{r,\beta}|$ negative bags, each containing a single image patch from the set $X^{r,\beta}$.

Output: Object bounding box A_t

2.2 Long-Term Detector

Object detection enables us to re-initialize the frame-to-frame tracker since it doesn't maintain an object model and unable to recover from failure. While the frame-to-frame tracker depends on the location of the object in the previous frame, the object detection mechanism presented here employs an exhaustive search in order to find the lost object.

Due to the efficiency of randomized ferns classifier [27] which is widely used in object recognition [2, 25, 26], we employ it as long-term detector to find possible object location. Ferns classifier consists of a number of ferns which are evaluated in parallel on each patch and fast. Each leaf in a fern records the number of positive p and negative n examples using Binary Pattern features during training. For a test sample, its evaluation by calculating the binary pattern features leads to a leaf in the fern. After that, the posterior probability for that input testing sample in feature vector x_i to

be labeled as an object ($y = 1$) by a fern j is computed as maximum likelihood estimator $\Pr_j(y = 1 | x_i) = p / (p + n)$, or is set zero if the leaf is empty. The final probability is calculated by averaging the posterior probabilities given by all ferns:

$$\Pr(y = 1 | x_i) = \sum_{j=1}^T \Pr_j(y = 1 | x_i) \quad (7)$$

where T is the number of ferns. Short-term trackers controls the posterior by adding its positive and negative samples to the ferns according to P-N constraints as [2] and multiple instance bag [3]. The P-constraints force all samples close to the validated trajectory to have positive label, while N-constraints have all patches far from the validated trajectory labeled as negative. Differently from [2], we bring in multiple instance bags around the validated trajectory so as to avoid the following problems. Slight inaccuracies in the tracker can therefore lead to incorrectly labeled training examples, which will further lead to the classifier resolving the ambiguities by itself to yield robust tracking results.

2.3 Samples Selection

A good classifier needs to have high prediction accuracy and generalization capability. The training samples' quality is crucial, especially for the training of online classifiers. In this paper, we introduce the bag samples selection to enhance the robustness of P-N constraints, which is able to use both weakly labeled and unlabeled bags.

The P-N constraints explore the latent information that there are some spatial structure and temporal structure information among different patches in video sequences. The constraints assume that a single object appears in one location only and therefore its trajectory defines a curve in the spatial-temporal volume. The trajectory curve is not continuous and generated by adaptive Lucas-Kanade [13] tracker and evaluated by the patch selected in the first frame using NCC measure to evaluate the confidence. P-constraints require that all patches that are close to validated trajectory have positive label. N-constraints require all patches in surrounding of a validated trajectory have negative label. In this paper, to mine and use the latent information effectively, especially to improve the generation capability of long-term detector, we sample the positive bag based on the patches close to validated trajectory and training online detector with the instance of the positive bag with soft-label. For detail, in our weakly supervised multiple instance learning tracker, training data has the form $\{(X_1, y_1), \dots, (X_n, y_n)\}$, where a bag $X_i = \{x_{i1}, \dots, x_{im}\}$ and y_i is a bag label. The bag labels are defined as:

$$y_i = \max_j (y_{ij}) \quad (8)$$

where y_{ij} are the instance labels, which are not known during the training. Since we assume the patches in or very close to validated trajectory as positive instance, the bag

which contains the patches is positive bag. The bag samples could be used for training the long-term detector.

2.4 Collaborative Training and Online Update

Frame-to-Frame tracker is used for motion estimation and collects the new templates which have high confidence with the old templates in the past validated trajectory of object appearance resized patches. It will be re-initialized by the final result fusing the trackers' and detector's result in the previous frame.

Adaptive weakly supervised tracker will be trained under the weakly supervised information coming from Frame-to-Frame tracker so that it could adapt to more cluttered background and prevent from drifting. Additionally, it could recommend more likely training samples for detector learning selection, especially in the case that detector hasn't been trained enough so as to fail to detect the possible candidates.

Learner will select the appropriate training data to train the long-term detector. For improving the detector's generalization capability, we generate multiple instance bags based on the predicted object location which is in the validated trajectory. For simplicity, we relax the condition of positive training examples and think that the instances' label is same to the bag's label:

$$y_{ij} = y_i \quad (9)$$

where y_{ij} is the label of the j^{th} instance in the i^{th} bag and y_i is the i^{th} bag's label. Additionally, the instances in one same bag should be satisfied that:

$$X_i = \{x : \|l(x) - l_i^*(x)\| < s\} \quad (10)$$

where X_i is the i^{th} bag, $l_i^*(x)$ is the predicted object location which is in the validated trajectory, $l(x)$ is the image patch's location, s is the bag's radius.

2.5 Result Fusion of Trackers and Detector

To fusing the results of the frame-to-frame tracker F_t , the weakly supervised tracker A and the confident detections D_t into a final result B_t is given. The decision is based on the number of detections, the detector's confidence values P_D^+ and the confidence of the tracking results P_R^+ , P_A^+ . The latter is obtained by running the template matching method on the tracking results. If the detector yields exactly one result with a confidence higher than the result from the trackers, then the response of the detector is assigned to the final result. The frame-to-frame tracker will be re-initialized by the final result. If the frame-to-frame tracker produced the most confident result, the result will be assigned to the final result. If their confidents are all high, we combine them by median selection. If P_R^+ and P_D^+ is low, we choose to believe the adaptive tracker. If P_A^+ is bigger than a threshold, the A is assigned to the

final result. In other cases the final result remains empty, which suggests that the object is not found in the current frame.

3 Experimental Results

In order to evaluate the performance of the proposed tracking approach, we test our system in C++ on several challenging image sequences. Nine videos (David, Jumping, Animal, Shaking, Cliffbar, Faceocc, Faceocc2, Surfer, Sylv) [10, 20, 3] are collected from the public dataset. The challenges of these videos include illumination variation, partial occlusion, pose variation, background clutter and scale change. For cross-validation, the center position error is compared with that of current state-of-the-art methods (FT[12], L1[29], MIL[3], and TLD[10]). We implemented these trackers using publicly available source code or binaries provided by the authors. They were initialized using their default parameters.

Table 1. Average center location error (pixels). The best performance is in bold, the second best is in underlined.

Sequence	#Frames	FT[12]	L1[29]	MIL[3]	TLD[10]	OURS
David	761	90	51.9	39.9	<u>14.9</u>	5
Jumping	313	58.2	50.8	12.6	<u>5.6</u>	4.7
Animal	71	91.2	160.5	<u>27.9</u>	86.6	12.1
Shaking	365	61.7	117.7	<u>51.4</u>	231.8	23.3
Cliffbar	328	17.7	43.3	13.8	50.7	<u>16.5</u>
Faceocc	886	5.7	<u>6.6</u>	35.3	11.3	13.7
Faceocc2	812	15.5	30.4	<u>12.2</u>	14.8	6.8
Surfer	376	139	37.7	<u>16.1</u>	18.1	15.9
Sylv	1344	<u>13.3</u>	34.5	14.7	9.4	<u>13.3</u>

The performance of visual trackers is evaluated according to the average per-frame distance (in pixels) between the center of the tracking result and that of ground truth. Clearly, this instance should be small. In Fig.2, we can see that our tracker consistently produce s a smaller distance than other trackers. This implies that our method can accurately track the target despite illumination changes.

At the same time, performance evaluation on public datasets is measured by Precision/Recall [2]. The results are displayed in Table 2.

The *David* sequence has large illumination changes. The initialized box makes many generative models fail in several frames. TLD and our method add the motion estimation information so as to prevent the target from missing. It's also very important for one appearance model updating. In the *Shaking* sequence, the tracked object is subject to changes in illumination and pose. TLD will fail in frame 58 because of abrupt powerful light. Our method will work because of the short-term tracker. When abrupt motion and large appearance changes simultaneously, our algorithm may fail.

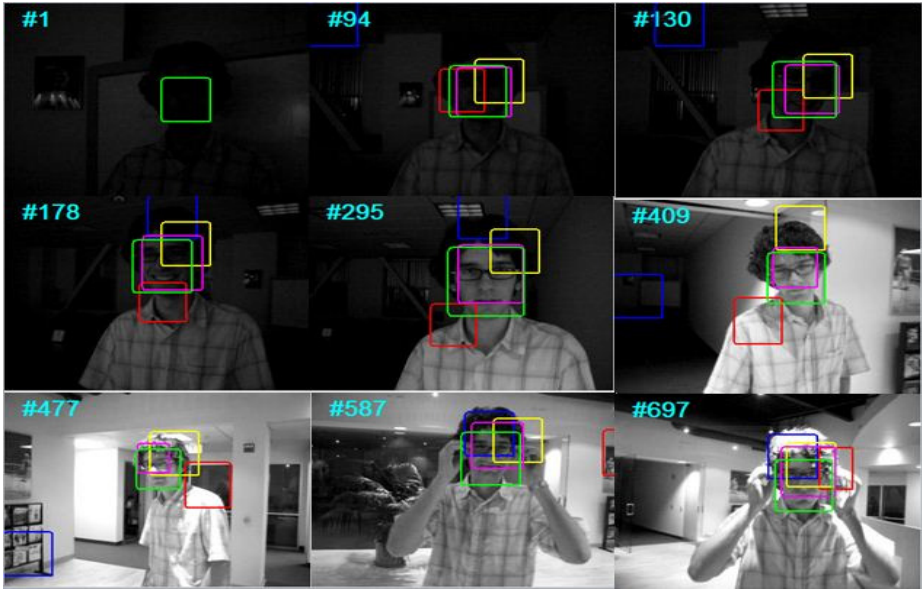


Fig. 2. Representative frames on sequences David under illumination changes. Blue, red, yellow, magenta and green bounding boxes were generated by FT, L1, MIL, TLD, and ours, respectively.

Table 2. Performance evaluation on public dataset measured by Precision/Recall. Bold numbers indicate the best score. The dataset is same as Table 1.

Sequence	FT[12]	L1[29]	MIL[3]	TLD[10]	OURS
David	0.158/0.158	0.309/0.309	0.143/0.143	0.999/0.999	1.000/1.000
Jumping	0.204/0.204	0.179/0.179	0.978/0.978	1.000/0.997	1.000/1.000
Animal	0.042/0.042	0.056/0.056	0.887/0.887	0.981/0.746	1.000/1.000
Shaking	0.397/0.397	0.063/0.063	0.825/0.825	1.000/0.156	0.893/ 0.893
Cliffbar	0.393/0.393	0.305/0.305	0.909/0.909	0.942/0.591	0.893/ 0.893
Faceocc	1.000/1.000	1.000/1.000	0.997/0.997	1.000/1.000	1.000/1.000
Faceocc2	1.000/1.000	0.702/1.000	1.000/1.000	1.000/1.000	0.974/0.974
Surfer	0.221/0.221	0.093/0.093	0.646/0.646	0.774/0.774	0.787/0.787
Sylv	0.885/0.885	0.467/0.467	0.858/0.858	0.949/0.949	0.955/0.955

The whole quantitative comparisons are shown in Table 1 and Table 2. From the tables, we can see that our tracking algorithm is better than the others in most cases.

4 Conclusion

In this paper, we propose a novel framework exploring their mutual relationship of adaptive trackers and detector and fusing them to act on visual tracking. Our method combines the flexibility of multiple instance learning on where to select positive

updates, the effectiveness of frame-to-frame tracking on object motion estimation and the robustness of detector towards partial occlusion and disappearance. In order to alleviate the drift of adaptive multiple instance tracker, we use the weakly supervised information coming from the frame-to-frame tracker. For improving the detector's generation capability, P-N constraints for bag samples selection are introduced to train the detector. Experimental results show the superiority of our approach over state-of-the art methods.

Acknowledgments. This work was supported by 973 Program (2010CB327905) and National Natural Science Foundation of China (61070104, 61273034, 61202325, 60905008).

References

1. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A Survey. *ACM Computing Surveys* 38(4) (2006)
2. Kalal, Z., Matas, J., Mikolajczyk, K.: P-N Learning: Bootstrapping Binary Classifiers by Structural Constraints. In: *Conference on Computer Vision and Pattern Recognition* (2010)
3. Babenko, B., Yang, M.H., Belongie, S.: Visual Tracking with Online Multiple Instance Learning. In: *Proc. CVPR* (2009)
4. Grabner, H., Bischof, H.: On-line boosting and vision. In: *CVPR* (2006)
5. Collins, R., Liu, Y., Leordeanu, M.: Online Selection of Discriminative Tracking Features. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27(10), 1631–1643 (2005)
6. Avidan, S.: Ensemble Tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(2), 261–271 (2007)
7. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised On-Line Boosting for Robust Tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part I. LNCS*, vol. 5302, pp. 234–247. Springer, Heidelberg (2008)
8. Tang, F., Brennan, S., Zhao, Q., Tao, H., Santa Cruz, U.C.: Co-tracking using semi-supervised support vector machines. In: *ICCV* (2007)
9. Yu, Q., Dinh, T.B., Medioni, G.G.: Online Tracking and Reacquisition Using Co-trained Generative and Discriminative Trackers. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part II. LNCS*, vol. 5303, pp. 678–691. Springer, Heidelberg (2008)
10. Kalal, Z., Mikolajczyk, K., Matas, J.: Tracking-Learning-Detection. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 6(1) (2010)
11. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: *CVPR* (2001)
12. Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: *CVPR* (2006)
13. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: *International Joint Conference on Artificial Intelligence*, vol. 81 (1981)
14. Shi, J., Tomasi, C.: Good features to track. In: *CVPR* (1994)
15. Matthew, I., Ishikawa, T., Baker, S.: The Template Update Problem. *IEEE TPAMI* (2004)
16. Black, M.J., Jepson, A.D.: Eigenttracking: Robust matching and tracking of articulated objects using a view-based representation. *IJCV* (1998)
17. Ross, D., Lim, J., Lin, R., Yang, M.: Incremental Learning for Robust Visual Tracking. *IJCV* (2007)

18. Wang, S., Lu, H., Yang, F., Yang, M.-H.: Superpixel Tracking. In: CVPR (2009)
19. Kwon, J., Lee, K.M.: Visual Tracking Decomposition. In: CVPR (2010)
20. Liu, B., Huang, J., Yang, L., Kulikowski, C.: Robust Tracking Using Local Sparse Appearance Model and K-Selection. In: CVPR (2011)
21. Kalal, Z., Mikolajczyk, K., Matas, J.: Forward-Backward Error: Automatic Detection of Tracking Failures. In: ICCV (2010)
22. Bradski, G., Kaehler, A.: Learning OpenCV: Computer Vision with the OpenCV Library, 1st edn. O'Reilly Media (2008)
23. Lewis, J.P.: Fast normalized cross-correlation. In: Vision Interface. In: Canadian Image Processing and Pattern Recognition Society (1995)
24. Dollar, P., Tu, Z., Tao, H., Belongie, S.: Feature Mining for Image Classification. In: Proc. IEEE Conf. CVPR (2007)
25. Bosch, A., Zisserman, A., Muoz, X.: Image classification using random forests and ferns. In: ICCV (2007)
26. Dinh, T.B., Vo, N., Medioni, G.: Context Tracker: Exploring Supporters and Distracter in Unconstrained Environments. In: CVPR (2011)
27. Ozuysal, M., Fua, P., Lepetit, V.: Fast keypoint recognition in ten lines of code. In: CVPR (2007)
28. Viola, P., Platt, J.C., Zhang, C.: Multiple Instance Boosting for Object Detection. In: Proc. Neural Information Processing Systems (2005)
29. Mei, X., Ling, H.: Robust Visual Tracking using L1 Minimization. In: ICCV (2009)

A Real-Time Fluid Rendering Method with Adaptive Surface Smoothing and Realistic Splash

Pengcheng Wang, Yong Zhang, Dehui Kong, and Baocai Yin

Beijing University of Technology
kdh@bjut.edu.cn

Abstract. We present an adaptive approach in particle-based fluid simulation to smooth the surface rendered using splatting in screen space. A real-time effect of surface smoothing and edge preserving is achieved in both the situations that camera is close to or far away from the fluid. This method is based on Bilateral filtering and using an adaptive range coefficient according to the viewing distance, so that the filter offers more blurring effect while the camera is approaching the surface and more edge protection when the viewpoint is maintaining a long distance to the fluid. We also introduce a physics-based splash model in turbulent flow for real-time simulation with a corresponding rendering method. The local density of particles in SPH simulation and Weber number are used to determine the formation and breakup of splash particles. Based on the splash breakup regime in physics, a pattern is proposed to organize the shape formed by the newly generated breaking up particles.

1 Introduction and Related Work

For the real-time representation of fluid, particle-based simulation method, such as Smoothed Particle Hydrodynamics (SPH) [1] [2], is usually preferred to Eulerian fluid representation, not only because particle-based representation (or Lagrangian representation) is simpler and faster, but also the method allows fluid to flow anywhere and interact with other objects in the scene. However, since the results obtained from particle-based simulation are discrete points, it is difficult to extract a surface for rendering.

In [3], a screen space point splatting method is proposed. This approach integrates screen space rendering presented in [4], which generates a surface only where it is visible, with a point splatting method to eliminate the generation of a mesh. Although this approach provides credible effect, real-time performance and inherent level-of-detail, there are still some apparent artifacts. Due to the spherical shape of the particle, it is an essential job to smooth the surface to make it look like real fluid. Cords et al. [3] use a Binomial filter which is able to smooth the surface but can't preserve the edges. A curvature flow filter is introduced in [5] to smooth the fluid generated in the same way, getting a relatively smoother surface, but needing to set an artificially controlled threshold

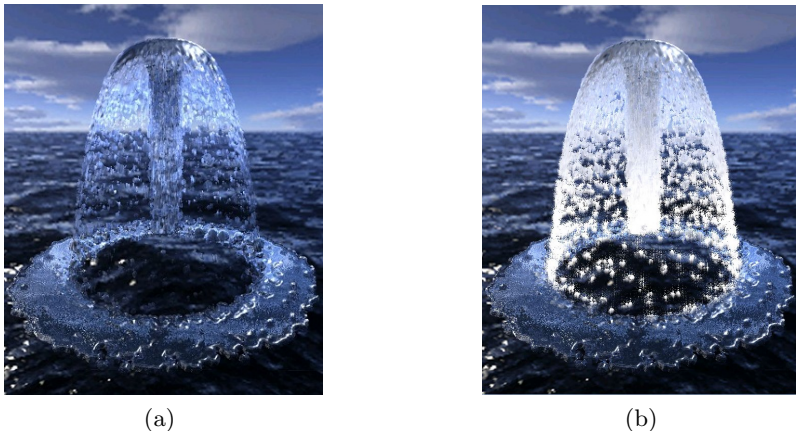


Fig. 1. (a) Fountain without splashes, (b) Fountain with splashes

to prevent blurring over silhouette. Bagar et al. [6] improve the curvature flow filter with an adaptive method to control the iteration times of blurring, leading to a proper smoothing extent in both near and far viewing distance, but still with the difficulty of setting an edge preserving threshold.

Some special effects of fluid, like splashes, bubbles and foam can significantly improve the appearance of turbulent fluid. Most of these elements have been seen in offline rendering method [12], and some of them have been successfully transplanted into real-time rendering [5] [6] [11]. Van der Laan et al. [5] bring in a foam effect to SPH fluid simulation using a grayish noise texture. A physics-based foam model proposed in [6] produces a beautiful underwater foam effect. However, these methods are not physically suitable for modeling splashes in a turbulent flow.

In this paper, we present a real-time rendering method for particle-based fluid simulation with the following advantages:

- We introduce an Adaptive Bilateral filter to obtain a smooth fluid surface with a good preservation of edges regardless of the viewing distance.
- We introduce a physics-based splash model using the local density and Weber number, getting a vivid splash breaking up effect in a turbulent flow.
- All the effects can be achieved in real-time with tens of thousands of particles in SPH simulation.
- The method is simple to be implemented directly on graphics hardware.

2 Adaptive Bilateral Filter

Our method is based on SPH fluid simulation which we assume has already offered us locations, velocities and densities of a set of particles. The rendering

approach we use is a screen space rendering method that has a similar process with it in [5], which can be generally summarized as following: First, we draw all the particles as point sprites into two separated textures, one for the depth of particles, from which the fluid surface can be extracted, and the other for the thickness of the fluid at each pixel, by using point splatting algorithm. Second, a blurring pass is executed in order to smooth the surface to prevent a bobbly appearance. Then, the normal at each pixel is calculated by reversing the coordinates in depth texture into view space. Finally, all the intermediate results are composited into a final one by appropriate illumination calculation.

In this process, surface smoothing is a vital step, because a spherical appearance for particles would significantly reduce the sense of reality. Some ordinarily used filters in Image Processing like Gaussian filter and Binomial filter will cause blurring over silhouette edges. In [5], the curvature flow method is used to adjust the depth value iteratively according to the curvature, meanwhile setting a threshold on depth difference when calculating the curvature to prevent from blurring edges. However, the confirmation of that special threshold makes it difficult to use, even for its adaptive edition in [6].

An obvious choice for not only smoothing but also edge preserving is using the Bilateral filter, like the one proposed in [7]. Although complained for non-separable and time-consuming, Bilateral filter offers a relatively good processing result. A spiral Bilateral filter in [7] is defined as follow:

$$h(a_0) = k^{-1} \sum_{i=0}^{n-1} f(a_i) \times g(a_i) \times r(a_i) \quad (1)$$

where $h(a_0)$ denotes the new depth value after filtering at pixel a_0 , $f(a_i)$ represents the original depth value in pixel a_i , $g(a_i)$ and $r(a_i)$ are the domain and range coefficient respectively which are determined by the distance and depth difference between pixel a_0 and a_i , and computed as follows:

$$g(x, y, t) = e^{-\frac{x^2+y^2}{2t}} \quad (2)$$

where t is a scale parameter, and

$$r(a_i) = e^{-\frac{[f(a_i)-f(a_0)]^2}{2\sigma_r^2}} \quad (3)$$

where σ_r^2 is the variance of depth value in the texture. The parameter k acts as a normalization constant and is defined as

$$k = \sum_{i=0}^{n-1} g(a_i) \times r(a_i) \quad (4)$$

However, after a careful observation on the two aims we want to achieve, a contradiction in Bilateral filter can be found. If a smoother fluid surface is intended, we need a bigger range coefficient $r(a_i)$ (in Equation 1). By contrary, if a sharper edge is wanted, a smaller range coefficient $r(a_i)$ approaching to 0

is better. In case we calculate $r(a_i)$ as in Equation 3, artifact caused by this contradiction that a jelly-like surface or blurring over silhouette edges can be observed apparently when the viewpoint is quite near or far from the fluid(see Figure 2 and Figure 3). This phenomenon can be explained as follows: After the perspective projection, those particles closer to the camera will have a relatively larger radius in screen space, while those further from the viewer will appear comparatively smaller on the screen, although they all have the same radius in a common reference coordinate system. So, with a range coefficient computed as in Equation 3, the same extent of blurring executed in the homogeneous clip space would bring an unqualified effect in screen space.

Therefore, we propose an approach to confirm the range coefficient $r(a_i)$ adaptively based on the viewing distance to the fluid, namely the depth value. For the same pixel a_0 on the screen, the larger depth value it has, the smaller range coefficient $r(a_i)$ it will obtain for its surrounding pixels; and the smaller depth value it has, the larger $r(a_i)$ it will have. If $r(a_i)$ increases to 1, the range coefficient would totally lose its impact, then our Adaptive Bilateral filter degenerates into Gaussian filter, which will bring the best blurring effect without any consideration of edge preserving. So, in our method, the range coefficient is calculated as follow:

$$r(a_i) = e^{-[f(a_i)-f(a_0)]^2 \times C(depth)} \quad (5)$$

Here we define $C(depth)$ as a piecewise function with a subsection point that is between 0 and 1.

$$C(depth) = \begin{cases} \sin(\frac{depth-k}{2k} \times \pi) + 1 & 0 < depth \leq k \\ B \cdot \sin(\frac{depth-k}{2(1-k)} \times \pi) + 1 & k < depth < 1 \end{cases} \quad (6)$$

In Equation 6, k determines whether more smoothing or more edge preserving is needed in a particular depth, and B is a constant larger than 1. When the depth value ranges from 0 to k , more smoothing is performed with $C(depth)$ returning a small value approaching to 0. If the depth value is larger than k , $C(depth)$ will return a large value approaching to $B + 1$ which preserves more unapparent edges. We choose to use a sin function because when the depth value is very small or very large, an extremely smoothing or edge preserving effect is wanted. It is not a transition that a simple function, like a linear function can offer. The constants k and B are confirmed by experiments.

In our method, we choose a fixed number of pixels as a_i in x and y direction in the range of the particle's diameter in screen space. Although it's not an efficient way considering data reading from the texture, the iteration step as in [5] could be eliminated. So, computation time caused by smoothing is not that much.

3 Real-Time Splash

A method of incorporating splashes into real-time fluid simulation and rendering is introduced in this section. We define splashes here as air bubbles trapping into

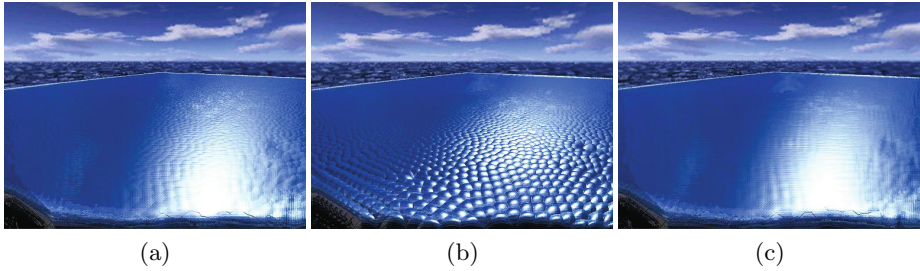


Fig. 2. Comparison between different filters in a near viewing distance. (a) Adaptive Bilateral filter, (b) Bilateral filter, (c) Gaussian filter.

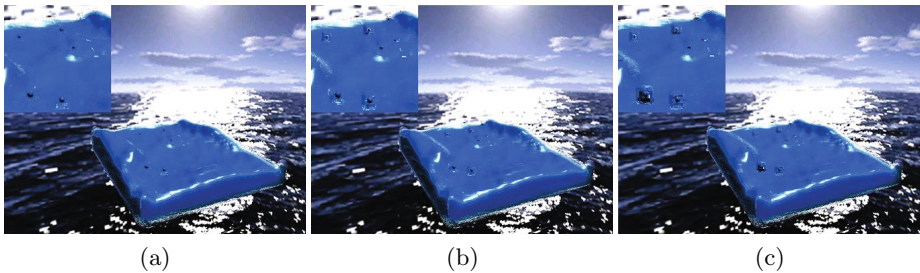


Fig. 3. Comparison between different filters in a far viewing distance. We place a close-up view of the separated particles on the top-left corner. (a) Adaptive Bilateral filter, (b) Bilateral filter, (c) Gaussian filter.

high-speed flow and causing fluid drops breaking up into smaller ones. This kind of phenomenon can often be observed in a fountain, waterfall or some turbulent flows.

3.1 Splash Formation and Breakup

First, we focus on how splashes form and break up. The two conditions we use here are particle's local density and Weber number [8]. In SPH algorithm, we can obtain a variable representing the local density of a particle. This variable reflects a regional quantity of particles near the center one. If a particle is surrounded by a large quantity of other fluid particles, it will get a high local density; while a particle only gets a low local density value when there are just a few or none fluid particles around it. So, having a low local density means a fluid particle can have access to much air, which is a fundamental condition for a water particle to turn into splash. In our method, we set a threshold on local density to distinguish those particles having possibilities to turn into splashes from those which are totally encircled by other fluid particles. We mark those particles as candidate splash particles for the final rendering, in which stage we will finally determine

whether these particles will be rendered as splashes or not. This is a direct and efficient way to do so.

After electing candidate splash particles, a more interesting phenomenon has to be paid attention to. We are aware that in a turbulent flow, those splashes cannot keep their shapes. In their moving process, they break up into small droplets when merging into the air. A lot of researches and experiments have been done on the splash (spray) breakup regimes, like in [9] and [10], which reveal that Weber number is a direct reflection of the splash foaming and breaking up occasions. Weber number, as defined in [8], is a dimensionless number in fluid mechanics as follow:

$$We = \frac{\rho v^2 l}{\sigma} \quad (7)$$

where ρ is the density of the fluid (this one is a constant describing the property of the fluid, not the same with the local density as described before), v is the relative velocity between the liquid and its surrounding air, l is the characteristic length which is taken as the particle's diameter in our simulation, and σ represents the surface tension of the fluid, which we assumes as a constant too. Weber number represents the ratio between inertial and surface tension forces. According to [10], fluid drop breakup occurs when its Weber number reaches to a threshold, as the kinetic energy of the liquid exceeds its surface energy to a certain extent. And the higher the Weber number is, the smaller the size and the larger the quantity of the breakup droplets will be. And from the experiment result in [10], with a certain Weber number, the shape of the breakup droplets looks like a comet with its tail pointing to the opposite direction to its velocity. Depending on these theories and experiments, we build such a splash model in SPH simulation. The entire splash breakup process will follow a pattern, which is an iterative process. If a candidate splash particle iterates once for breaking up, eight smaller splash particles will be generated. These eight particles will be located around the original one, and arranged into a comet shape as in Figure 4:

To get a comet shape, four particles are generated in the direction along or perpendicular to the original particle's velocity direction. The other four particles are distributed symmetrically in between the previous four. The distances from the new particles to P_0 are arranged from R_a to R_c and R_d to R_e in ascending order. R_a to R_e would shrink in a certain extent after one iteration to keep the new particles generated in the next iteration encircled in the comet shape this time, if there are any. We just consider the velocity of P_0 in x and y coordinates in screen space to keep the comet plane always facing to the camera.

In order to determine how many times this breakup process would iterate for, we define the following equation:

$$T = \lfloor We/C \rfloor \quad (8)$$

in which an empirical constant C is defined as a criterion. C should not be set to a exceedingly small constant, for too many new particles will tremendously

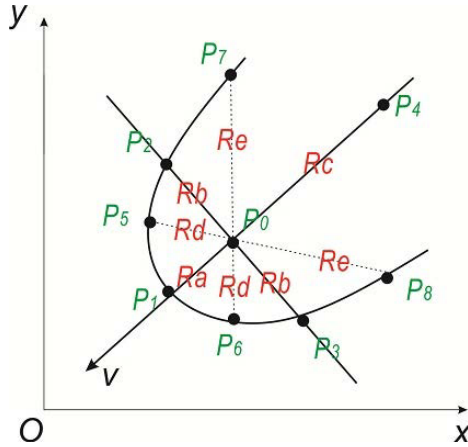


Fig. 4. The shape arranged by the newly generated particles in splash breaking up for one iteration. P_0 represents the original particle, P_1 to P_8 are the particles generated after breakup, R_a to R_e are their respective distance to P_0 . The velocity direction of P_0 is marked by v .

affect the rendering efficiency. To determine the size of the new particles after breakup, we use the following equation:

$$R = R_0 / \text{sqrt}(We) \tag{9}$$

where R_0 is the original radius of P_0 , We is P_0 's Weber number. Attention should be paid to those particles with a small Weber number under 1.0, because it will have a deviant large radius after the computation in Equation 9. So we should guarantee the denominator in Equation 9 is in a rational range by some additional examination. Other properties of the newly generated particles should be the same as those of P_0 before breakup.

So, in a conclusion, our splash formation and breakup model works in the following steps:

- To all the particles in a SPH simulation, we set a threshold of the local density to label those particles who have a chance to contact with surrounding air and turn into splashes as candidate splash particles.
- Calculate the Weber number of those candidate particles. Determine the iteration times of the breaking up process, the distance to the original particle and the size of the newly generated particle according to the Weber number.
- Generate eight new particles for one iteration step in the comet shape, set all the properties of the new particles and finish all the rest iterations.

3.2 Splash Rendering

In our approach, we choose to not consider the splash forming and breaking up in SPH simulation, but only in the rendering steps. Some reasons here can support

our decision. First, splash particles after breaking up will have a relatively small quality and size, which means they will have a very little influence on other particles. If we ignore those influences, no obvious artifacts can be observed. Second, if we just shrink the radius of the splash particles in rendering process, but not in SPH simulation, the particle P_0 will have the same influence on other particles as before the breaking up, which makes it look like having much more influence compared with the size of the shrunk particle after breakup. That could be a visual compensation for the new particles' impact in interactions among the fluid. Also, SPH simulation can't handle an arbitrarily large amount of particles if we intend to run the simulation in real-time. Sometimes we could generate up to 24 new particles from one candidate splash particle. And adding new particles occasionally to a SPH simulation is difficult and not inefficient. So, we just do all the work in rendering process.

Our splash breakup method can be naturally mapped to geometry shader operation. We examine the local density of each particle, compute the Weber number of each candidate particle, and break them up into new splash particles totally in geometry shader. These processes should be done in the depth and thickness computation passes in rendering.

Suppose that we have obtained the fluid surface rendering result without a splash effect, by integrating those intermediate results in depth, thickness, blurring and normal computing passes into a final one with Blinn-Phong shading model and Fresnel equation. If splash particles are on top of the fluid surface, which means Weber number is up to an extent there, we will mix the splash color into the fluid surface color. We can notice that the higher Weber number a particle has, the more new particles it will break into, which means the thicker the splashes will be there. So according to this conclusion, the final color of the fluid surface with splash should be determined by

$$C = C_f \times (1 - coe) + C_s \times coe \quad (10)$$

$$coe = We/N \quad (11)$$

where C_f is the color of fluid without splashes, C_s is the color of splash, coe is a coefficient between 0 and 1, determined by Weber number We and an empirical constant N . To increase the sense of reality, we also use motion blur as a post-processing step, which can offer a better effect of the turbulent flow.

4 Results

We have tested our approach in two scenes. In the water pool scene, a giant water ball falls into a pool beneath it. Our Adaptive Bilateral filter is tested in this scene, where a peaceful water surface and isolated particles are both easy to be observed. The fountain is an excellent scene to test our splash formation, breakup and rendering method. All benchmarks were performed on an NVidia Quadro FX 5800 graphics card with 240 shader cores and 4G graphics memory. The SPH simulation is done by NVidia PhysX 2.8.1. In all the computation time listed below, the simulation time is not included.

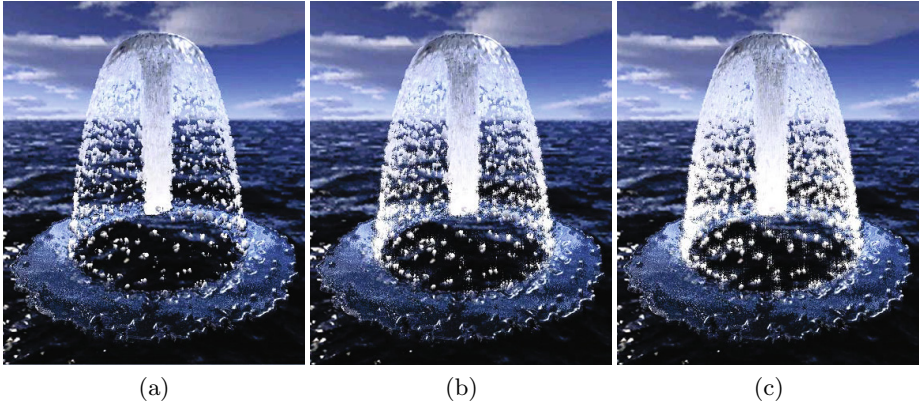


Fig. 5. Comparison of the effect with different splash breakup iteration times. (a) No splash breakup, (b) Half of the iteration times, (c) Full breakup iterations.

Table 1. Performance comparison between different smoothing method

Filter	Rendering Time (33371 particles) (in ms)	Rendering Time (65517 particles) (in ms)
Adaptive Bilateral	26.81	42.25
Bilateral	26.53	42.00
Gaussian	26.53	41.52
None	25.77	40.00

Table 1 shows the respective rendering time using our blurring method, Bilateral filter and Gaussian filter in the water pool scene. Figure 2 and Figure 3 show the comparison of the effects in this simulation. According to this group of contrast, we can see our smoothing method offers a preferable effect compared to Bilateral filter, whose effect is very close to the smoothed imaged blurred by Gaussian filter in a very near viewing distance. Meanwhile, the edges between the front water surface and back water surface are better preserved compared to Bilateral filter in a far view distance in our approach, and of course better than them with Gaussian filter due to Gaussian filter's ignoring of edge protection. The increased computing time in our method is acceptable, only 1.06% longer than it with Bilateral or Gaussian filter and 4.04% longer than it without any blurring in a 33371 particles scene. In a 65517 particles scene, the rendering time in our method is 0.60% longer than it with Bilateral filter, 1.81% longer than it with Gaussian filter, and 5.63% longer than it with none of filters.

In the fountain scene, our splash formation and breakup method is tested. Coming with the great improvement on appearance is a relatively big growth in the rendering time. The total amount of particles is 65000 and the rate of

Table 2. Performance comparison between rendering of the fountain with or without splashes

Splashes	Rendering Time (65000 particles) (in ms)
With Splashes	16.63
Without Splashes	22.27
Screen Full of Splashes	26.25

emitting is 6000 particles per second in this scene. As revealed in Table 2, compared to the method without a splash effect, our method’s rendering time is 33.91% longer. When the screen is full of splash particles with a near viewpoint to the fountain, like in Figure 6, the rendering time is especially long, because the newly generated vertexes covers more pixels in this kind of situation. However, the reinforce of the rendering effect due to splashes is worth of the computation cost, like showed in Figure 1 and Figure 5.

**Fig. 6.** A close-up view of the splashes. Along with the velocity increment from the top down, the dispersing of the splashes becomes more apparent.

5 Conclusions and Future Work

In this paper, we present an Adaptive Bilateral filter which controls the extent of surface smoothing based on the viewing distance to the fluid. This method can lead to a smoother surface when the camera is close to the fluid, and preserve the edges at a far viewing distance. The second contribution of this paper is a physics-based splash model in real-time rendering of particle-based fluid. A splash breakup regime in aerodynamics is used to control the formation, breakup

occasion and shape formed by newly generated splash particles, which will increase the sense of reality of the rendering significantly.

Future work may involve looking at the method of modeling foam under water and integrating it into our splash model. Rendering of splashes behind other fluid surfaces efficiently is also an effect we want to achieve.

Acknowledgment. This research is supported by NSFC (No. U0935004, 60825203, 61100130).

References

1. Desbrun, M., Gascuel, M.-P.: Smoothed particles: A new paradigm for animating highly deformable bodies. In: Boulic, R., Hegron, G. (eds.) Eurographics Workshop on Computer Animation and Simulation (EGCAS), pp. 61–76. Springer, Heidelberg (1996)
2. Müller, M., Charypar, D., Gross, M.: Particle-based fluid simulation for interactive applications. In: Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA 2003, New York, USA, pp. 154–159 (2003)
3. Cords, H., Staadt, O.: Instant liquids. Poster Proceedings of ACM Siggraph/Eurographics Symposium on Computer Animation (2008)
4. Müller, M., Schirm, S., Duthaler, S.: Screen space meshes. In: Proceedings of the 2007 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA 2007, pp. 9–15. Eurographics Association (2008)
5. Van Der Laan, W.J., Green, S., Sainz, M.: Screen space fluid rendering with curvature flow. In: Proceedings of I3D 2009: The 2009 ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games, Boston, MA, United states, pp. 91–98 (2009)
6. Bagar, F., Scherzer, D., Wimmer, M.: A layered particle-based fluid model for real-time rendering of water. *Computer Graphics Forum* 29(4), 1383–1389 (2010)
7. Hu, Q., He, X., Zhou, J.: Multi-scale edge detection with bilateral filtering in spiral architecture. In: Proceedings of the Pan-Sydney area workshop on Visual information processing. CRPIT, vol. 36, pp. 29–32. Australian Computer Society (2003)
8. Weast, R., Lide, D., Astle, M., Beyer, W.: CRC Handbook of Chemistry and Physics, 70th edn. CRC Press (1990)
9. Borisov, A., Gel'fand, B., Natanzon, M., Kossov, O.: Droplet breakup regimes and criteria for their existence. *Journal of Engineering Physics* 40(1), 44–49 (1981)
10. Joseph, D., Belanger, J., Beavers, G.: Breakup of a liquid drop suddenly exposed to a high-speed airstream. *International Journal of Multiphase Flow* 25(6-7), 1263–1303 (1999)
11. Mihalef, V., Metaxas, D., Sussman, M.: Simulation of two-phase flow with sub-scale droplet and bubble effects. *Computer Graphics Forum* 28(2), 229–238 (2009)
12. Takahashi, T., Fujii, H., Kunimatsu, A., Hiwada, K., Saito, T., Tanaka, K., Ueki, H.: Realistic animation of fluid with splash and foam. *Computer Graphics Forum* 22(3), 391–400 (2003)
13. Zhang, Y., Solenthaler, B., Pajarola, R.: Adaptive sampling and rendering of fluids on the gpu. In: Proc. of Symposium on Point-Based Graphics, pp. 137–146 (2008)

Multi-document Summarization Exploiting Semantic Analysis Based on Tag Cluster

Jee-Uk Heu¹, Jin-Woo Jeong¹, Iqbal Qasim¹, Young-Do Joo²,
Joon-Myun Cho³, and Dong-Ho Lee¹

¹ Department of Computer Science and Engineering, Hanyang University,
Ansan, Kyeonggi-do, Korea

{hyugar, selphyr, qasim, dhlee72}@hanyang.ac.kr

² Division of Computer Media Engineering, Kangnam University
Yongin Kyeonggi-do, Korea

ydjoo@kangnam.ac.kr

³ SmartTV Research Center

ETRI(Electronics and Telecommunications Research Insitute), Korea

jmcho@etri.re.kr

Abstract. Multi-document summarization techniques aim to reduce the documents into a small set of words or paragraphs that convey the main meaning of the original documents. Many approaches for multi-document summarization have used probability based methods and machine learning techniques to summarize multiple documents sharing a common topic at the same time. However, these techniques fail to semantically analyze proper nouns and newly-coined words because most of them depend on old-fashioned dictionary or thesaurus. To overcome these drawbacks, we propose a novel multi-document summarization technique which employs the tag cluster on Flickr, a kind of folksonomy systems, for detecting key sentences from multiple documents. We first create a word frequency table for analyzing the semantics and contribution of words by using HITS algorithm. Then, by exploiting tag clusters, we analyze the semantic relationship between words in the word frequency table. The experimental results on TAC 2008, 2009 data sets demonstrate the improvement of our proposed framework over existing summarization systems.

Keywords: Multi-Document Summarization, Tag Cluster, Semantic Analysis.

1 Introduction

Recently, the rapid growth of Internet and smart multimedia devices such as smart phones and tablet PCs makes users to find information easily through diverse media (e.g., document, image, video, and music) from the Web. Especially, in Web 2.0 environment, users can take more general and common information from the folksonomy system where general users can add tags to describe the contents of multimedia. The most important and distinctive feature of the folksonomy system (e.g., Wikipedia, Flickr, del.icio.us) is the creation of contents by general users without any restriction. Users are now able to create, share, and search multimedia contents anytime and anywhere by using their smart devices. As a result, the amount

of documents (news, blog, Web page, email, etc) created on the Web has been rapidly increasing day by day. In these environments, in order to find the necessary information, users have to manually review all of the searched documents without any assistance of search engines, and it requires too much time and effort. To address this problem, various document summarization techniques have been studied to efficiently summarize the core of single original document. More recently, multi-document summarization techniques have been researched to summarize multiple documents sharing a common topic at the same time.

Most existing multi-document summarization techniques analyze the semantic relationships between the words in the documents by exploiting probability theory, machine learning techniques, and external knowledge-bases such as WordNet. However, these techniques suffer from high computational cost in learning and summarization processes. Furthermore, WordNet-based approaches fail to analyze proper nouns (e.g., person's name, product, firm name) and newly-coined words due to the absence of these words in WordNet[7].

In this paper, we propose a novel multi-document summarization technique using tag clusters of the folksonomy system to detect key sentences in multiple documents. For this purpose, we exploit the tag cluster serviced by Flickr, one of the most representative folksonomy systems, for summarizing multiple documents. Using tag clusters, we analyze the importance of each word and the semantic relatedness among them, and finally make a summary.

The rest of this paper is organized as follows: In Section 2, we briefly review the related works. The proposed multi-document summarization system is presented in Section 3. In Section 4, we describe the experimental results using TAC 2008, 2009 dataset. Finally, we conclude our work in Section 5.

2 Related Work

Multi-document summarization techniques can be classified into two approaches. One is extractive summarization approach and the other is abstractive summarization approach [1]. Extractive summarization approach involves assigning saliency scores to some units (e.g. sentences, paragraphs) of the documents and extracting those units with the highest scores. In contrast, abstractive summarization approach takes the essence of the source document to build a summary by using natural language processing techniques.

Although the abstraction-based method can summarize a document more accurately than the extraction-based method does, it is much more difficult and complex than extraction-based summary because it requires the use of high-costly natural language processing technologies such as information fusion [2], sentence compression [3], and reformulation [4].

Hennig et al. [5] proposed a multi-document summarization method based on Probabilistic Latent Semantic Analysis (PLSA), which represents sentences and queries as probability distributions over latent topics. They combine query-focused features and thematic sentence features into an overall sentence score. Xiaojun [6] proposed two novel summarization models to make use of theme cluster in the document set. The first model incorporates the cluster information in conditional

markov random walk model and the second model uses Hypertext Induced Topic Search (*HITS*) algorithm. Although these models had shown good performance, however, they need a high cost and more time when performing the clustering algorithm in the preprocessing step, and also employ the probability method which requires a complex computation.

Chenghua et al. [7] proposed a method to detect key sentences by using the keyword extraction based on statistics and synsets. They used WordNet for extracting the most relevant sentences from the original document. However, it is not easy to analyze the relationship between the semantics of the words in Web documents because there are many proper nouns and new words in the original documents which are not defined in WordNet.

Zhu et al. [8] proposed a tag-oriented Web document summarization approach by using both document itself and the tags annotated on the document. This approach has limitations on the semantic analysis between words in the document because they only analyze the relationships between users and tags in the folksonomy system.

To overcome these drawbacks, we propose a novel multi-document summarization technique using tag clusters of the folksonomy system to detect the key sentences in multiple documents. For this purpose, we exploit the tag cluster serviced by Flickr, one of the most representative folksonomy systems, for summarizing multiple documents. Using tag clusters, we analyze the importance of each word and semantic relatedness among them, and finally make a summary.

3 The Proposed System

Fig. 1 shows the framework of our multi-document summarization system. Given multiple documents that need to be summarized, first, we perform a pre-processing step so that the documents can be analyzed at different granularities (i.e., word-level and sentence-level). Afterwards, the words analysis module calculates the word frequency and analyzes the semantic importance of each word by exploiting tag clusters from Flickr. Then we compute the contribution of each word using *HITS* algorithm. The computed semantic importance score and contribution are used for rating each sentence in multiple documents. Finally, the sentence analysis module generates the final summary of multiple documents by selecting top-k ranked sentences. The main phase of the framework will be described in the following subsection.

3.1 Preprocessing

The preprocessing module extracts sentences from the input documents, and then performs tokenization and stopword elimination. A set of words generated after preprocessing are used by the word analysis module to compute the semantic importance and contribution of each word in the documents. Also, the extracted sentences will be given a weighted score based on the importance and contribution of the words in each sentence.

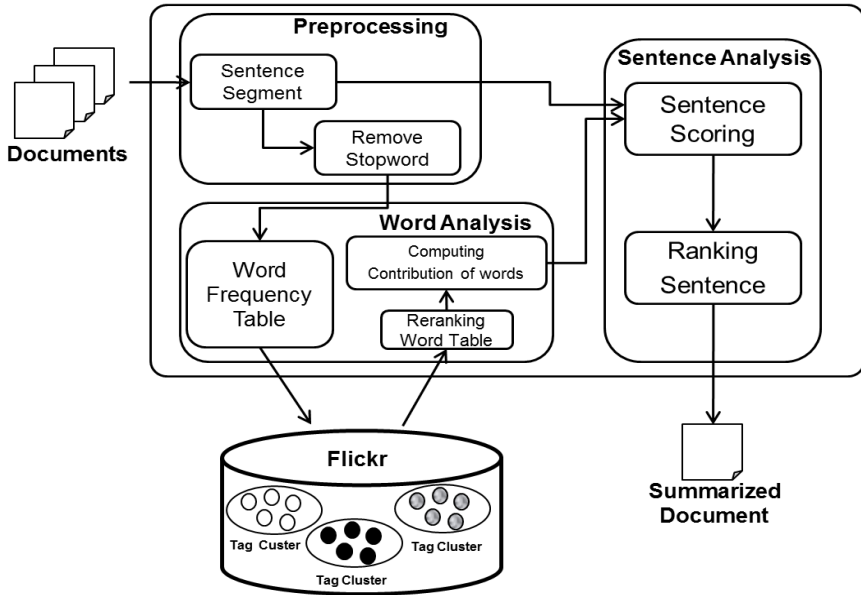


Fig. 1. System Overview

3.2 Word Analysis

To analyze how much each word contributes to the document, we construct Word Frequency Table (WFT) using tag clusters and exploit HITS algorithm.

Creation of Word Frequency Table. First of all, we construct Word Frequency Table to calculate the frequency of each word in the documents, which can be represented as:

$$WFT = \{(w_1, c_1), (w_2, c_2), \dots (w_{n-1}, c_{n-1}), (w_n, c_n)\}^T \tag{1}$$

In (1), $w_i = (w_1, w_2, \dots, w_n)$ is i -th word in documents, and $c_i = (c_1, c_2, \dots, c_n)$ is the frequency of i -th word. After the construction of WFT, it is sorted by the frequency c . However, it is hard to assertively say that the word with high frequency is surely important in the documents. Therefore, in addition to the frequency, we exploit the tag cluster that can be obtained from Flickr, which is one of the folksonomy systems, to analyze the semantics of words. A folksonomy is a system of classification derived from the practice and method of collaboratively creating and managing tags to annotate by user. The tag cluster in Flickr is a list of tags considered semantically similar in a tag space, so that tag clusters can provide useful information for semantic analysis of words. Also, one additional advantage of using Flickr is the coverage of a tag space. The tags from Flickr can cover most of the words such as proper nouns and newly-coined words that are not defined in WordNet.

Fig. 2 shows the procedure to construct the final WFT. Given the initial word frequency table, we obtain tag clusters of each word in WFT from Flickr. For

example, a tag cluster of word ‘airbus’ contains social tags such as ‘airport’, ‘plane’, and ‘a380’. It means that ‘airport’, ‘plane’ and ‘a380’ have high semantic closeness with ‘airbus’ in the tag space. Then, we collect every tag that corresponds with each word existing in original *WFT* and count the frequency of the collected tags to update the frequency of each word in *WFT*.

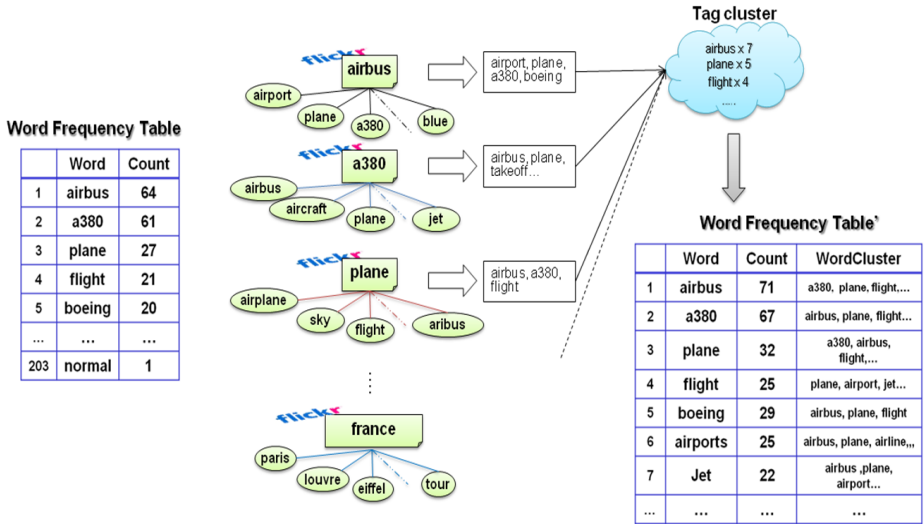


Fig. 2. A procedure for constructing *WFT'* using tag cluster from Flickr

The tags in each tag cluster are also stored together in *WFT*. The updated word frequency table (*WFT'*) in Fig. 2 shows the final table generated by this procedure, in which the count and WordCluster information are updated. The final word frequency table *WFT'* is represented as follows:

$$WFT' = \{(w_1, c_1, wc_1), (w_2, c_2, wc_2), \dots, (w_n, c_n, wc_n)\}^T \tag{2}$$

where wc_i is the WordCluster of w_i . *WFT'* is used for analyzing the contribution of each words in the documents in Section 3.3.

Analyzing Contribution of Word by Using HITS. For calculating the contribution of each word in *WFT'*, we apply *HITS* algorithm to our system [9]. Originally, the *HITS* is a kind of link analysis algorithm to rate Web pages. The *HITS* algorithm classifies each Web page as a hub and authority. A good hub represents a page that pointed to many good authorities, and a good authority represents a page that was linked by many good hubs. For analyzing the contribution of each word, we consider the WordCluster as an authority, and the words in *WFT'* as a hub. The *HITS* calculates a contribution score for every word in *WFT'* as follows:

$$HITS(w_i) = a(w_i) + h(w_i). \tag{3}$$

Let an authority score of w_i be $a(w_i)$ and a hub score of w_i be $h(w_i)$. If w_j in WFT' exists in WordCluster of w_i , authority score $a(w_i)$ increase by 1, otherwise gets zero. In contrast, if w_j in WordCluster of w_i exists in WFT' , then hub score $h(w_i)$ increases by 1, otherwise gets zero. Finally, the *HITS* score is calculated by summing an authority score and a hub score.

3.3 Sentences Analysis

After analyzing the contribution of each word, we calculate the sentence score and rank each sentence with WordCluster. Algorithm 1 shows the procedure for computing a sentence score. The importance of a sentence is determined by its corresponding score. In Algorithm 1, we define rel-gram as a contiguous sequence of k words without duplicates.

Table 1. Difference between n-gram and rel-gram

k	n-gram	rel-gram
1	A, B, C	A, B, C
2	AB, BC, BA, BC, CA, CB	AB, BC, AC
3	ABC, ACB, BAC, BCA, CAB, CBA	ABC

Table 1. shows the difference between N-gram and rel-gram. N-gram does consider the sequence of words, thus when $k=2$, 'AB' and 'BA' are different from each other. However, in case of rel-gram, they are same 2-gram because real-gram does not consider the sequence of words.

```

Algorithm1. SentenceScoring
1 for(  $n = 1; n < k; n++$  )
2   if(  $n == 1$  )
3     take one word  $w$  in the  $WFT'$ 
4     for each sentence  $s$ 
5       find  $s$  include  $w$ 
6      $ScoreTable \leftarrow Score(s, w)$ 
7     end for
8   else
9     compute rel-gram in  $WFT'$ 
10    for each sentence  $s$ 
11      find  $s$  include rel-gram
12     $ScoreTable \leftarrow Score(s, rel - gram)$ 
13    end for
14  end for
15   $Sort(ScoreTable)$  by SentenceScore
    
```

We assume that S is a set of sentences in the documents. First of all, Algorithm 1 calculates the scores of sentences that include single words (line 2-7). Afterwards, it calculates the scores of sentences that contain k rel-grams (line 9-13). The sentences are stored in Score Table along with their sentence scores (line 6 and line 12). These

sentences are sorted by their scores for selecting the sentences which more contribute to summarize multiple documents(line 15). The most remarkable feature of our sentence scoring method is to consider the semantic relationship among words in a sentence.

Fig. 3 shows an illustrative example of computing the relationship between words. Assume that there are three words ‘A’, ‘B’, ‘C’, and each word has its own WordCluster ‘Cluster_A’, ‘Cluster_B’, ‘Cluster_C’. As shown in Fig. 3, Cluster_A includes word ‘B’, but does not include word ‘C’. And Cluster_B includes both words ‘A’ and ‘B’ while Cluster_C has no words.

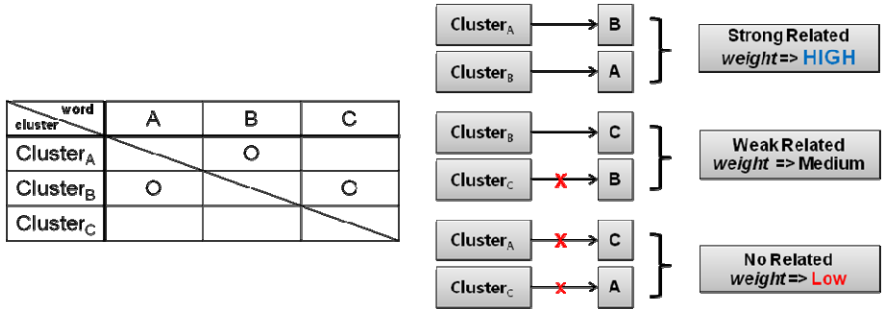


Fig. 3. A procedure for computing the semantic relationship between words

In this case, word ‘A’ and ‘B’ have high semantic relationship, because their WordCluster include both words. But word ‘B’ and ‘C’ have medium semantic relationship, because only Cluster_B includes ‘C’. Word ‘A’ and ‘C’ have no relationship because their WordCluster do not include each other word, respectively. We calculate the semantic relationship between rel-gram as follows:

$$Rel(rel - gram) = \sum_{i,j \in WFT'} Relate_{i,j} \tag{4}$$

where

$$Relate_{i,j} = \begin{cases} 1 & \text{if } w_j \in Cluster_{w_i} \\ 0 & \text{Otherwise} \end{cases} \tag{5}$$

And we compute the score of a sentence as follows:

$$Score(s, rel - gram) = \{\alpha \cdot \sum_{i \in WFT'} Freq(w_i)\} + \{\beta \cdot \sum_{i \in WFT'} HITS(w_i)\} + \gamma \cdot Rel(rel - gram) \tag{6}$$

where *s* is the sentence which includes the rel-gram. And *Freq*(*w_i*) is the frequency of *w_i*. *HITS*(*w_i*) is a score of contribution for each word. And *Rel*(*rel-gram*) is the score of the semantic relationship between rel-grams. Also, ‘α’, ‘β’, ‘γ’ are the weights of the each term (where α+β+γ =1). Finally, our system generates the final summary based on the top k scored sentences in Score Table.

4 Experiments

4.1 Data Set and Evaluation Metric

We use TAC 2008 and TAC 2009 data sets to test our proposed method empirically. Both data sets are open benchmark data sets from Text Analysis Conference (TAC) for automatic summarization evaluation. TAC 2008 provides 48 document sets and TAC 2009 provides 44 document sets. And we used ROUGE [10] toolkit for evaluations, which has been widely adopted by Document Understanding Conference(DUC) for automatic summarization evaluation. It measures the quality of document summarization by counting overlapping units such as n-gram, word sequences, and word pairs between the candidate summary (a summary by summarization techniques) and the reference summary (a summary by human experts). Several automatic evaluation methods are implemented in ROUGE, such as ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-SU. ROUGE-N is an n-gram recall measure computed as follows:

$$ROUGE - N = \frac{\sum_{S \in \{Ref\}} \sum_{n-gram \in S} Count_{match}(n-gram)}{\sum_{S \in \{Ref\}} \sum_{n-gram \in S} Count(n-gram)} \tag{7}$$

In (7), n is the length of the n-gram, and Ref stands for the reference summaries, $Count_{match}(n-gram)$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. $Count(n-gram)$ is the number of n-grams in the reference summaries. We show ROUGE-SU(skip bigram plus unigram) metrics in the experimental results.

4.2 Experimental Results

In our first experiment, we varied the weight α, β, γ of (6) from 0 to 1, and Fig. 4 and Fig.5 show the F-Measure of ROUGE-SU4 on TAC 2008 and TAC 2009 datasets, respectively. On TAC 2008, generally, our proposed system shows good performance

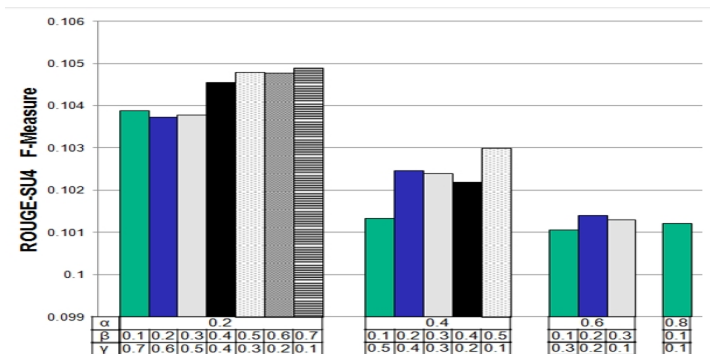


Fig. 4. Study of weight α, β, γ using ROUGE-SU4 on TAC 2008

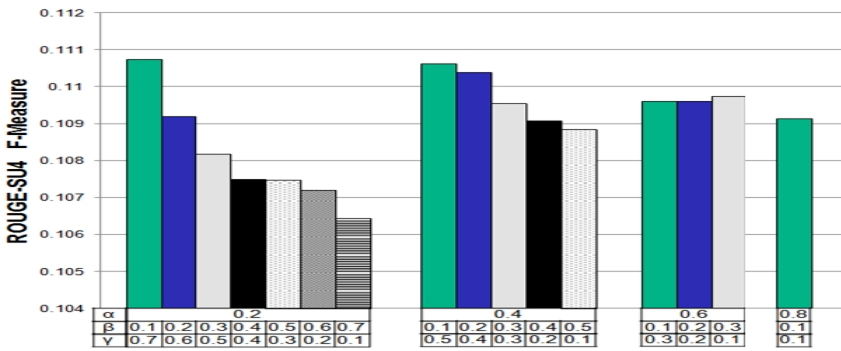


Fig. 5. Study of weight α, β, γ using ROUGE-SU4 on TAC 2009

regardless of β and γ when $\alpha = 0.2$. We can see that as the value of α increases, the performance gets worse. It indicates that the frequency of the word mainly affects the performance of summarizing documents. However, on TAC 2009, the performance of our proposed approach is mainly influenced by β and γ . This indicates that the quality of document summarization is mainly influenced by the contribution of words and semantic relationship between words. From our analysis, we found that the reference summary of each data set (i.e., TAC 2008 and TAC 2009) has different characteristics, so that the experimental results show different patterns.

In the second experiment, we set the weight α, β, γ to the values that showed the best performance in the first experiment (TAC 2008: $\alpha=0.2, \beta=0.7, \gamma=0.1$ and TAC 2009: $\alpha=0.2, \beta=0.1, \gamma=0.7$). Then we varied the value of k of rel-gram from 3 to 5. We also varied the number of words where rel-gram is computed from 10 to 20. Fig. 6 and Fig. 7 demonstrate the influence of summarization by changing the number of rel-gram. On TAC 2008, our proposed system shows good performance when $k = 3$, while shows good performance when $k = 4$ on TAC2009.

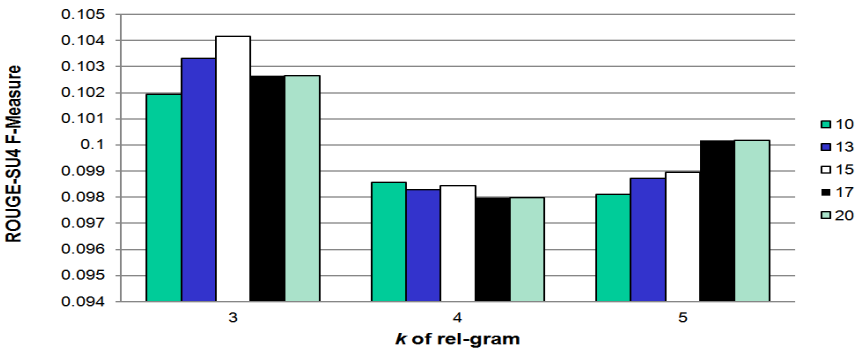


Fig. 6. Study of *rel-gram* words ROUGE-SU4 on TAC 2008

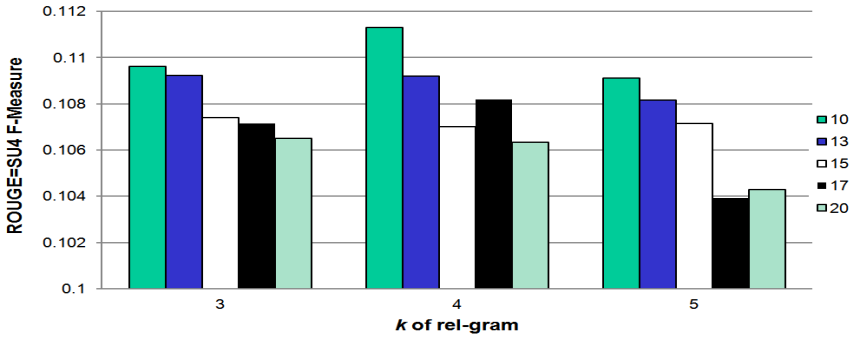


Fig. 7. Study of rel-gram words ROUGE-SU4 on TAC 2009

Table 2 and Table 3 show the results when comparing our proposed system with related techniques using ROUGE-2 and ROUGE-SU4 on TAC2008 and TAC2009, respectively. In this experiment, the system NIST, ceaList1, LPN1 and Veness Team1 are used as baselines. These result of other system is provided by TAC2008 and TAC2009. As we can see from Table 2 and Table 3, our proposed system outperforms other baseline systems on both TAC 2008 and TAC 2009 datasets.

Table 2. Comparison results on TAC 2008

	ROUGE-2			ROUGE-SU4		
	Recall	Precision	F-Measure	Recall	Precision	F-Measure
Our-System	0.0654	0.07383	0.06907	0.09935	0.11207	0.1049
System – NIST (Baseline)	0.05871	0.06870	0.06244	0.09300	0.10644	0.09822
System - ceaList1	0.06968	0.06976	0.06969	0.10683	0.10686	0.10679
System-LIPN1	0.03343	0.05017	0.03922	0.06498	0.09968	0.07680
System - VensesTeam1	0.06098	0.06517	0.06283	0.09805	0.10437	0.10084

Table 3. Comparison results on TAC 2009

	ROUGE-2			ROUGE-SU4		
	Recall	Precision	F-Measure	Recall	Precision	F-Measure
Our-System	0.07051	0.08223	0.07549	0.103976	0.12144	0.11131
System – NIST (Baseline)	0.05142	0.05861	0.05457	0.09091	0.10309	0.09624
System - ceaList1	0.05751	0.058795	0.0581	0.09622	0.09866	0.09735
System-LIPN1	0.05252	0.04977	0.05108	0.09654	0.09156	0.09393
System - VensesTeam1	0.07002	0.06503	0.06740	0.11617.	0.10772	0.11172

5 Conclusions

In this paper, we proposed a novel multi-document summarization technique using tag clusters of the folksonomy system to detect the key sentences in multiple documents. Our proposed system exploits Flickr to acquire tag clusters for analyzing the semantics of words. It efficiently summarizes the documents by detecting meaningful words and their semantic relatedness. The remarkable advantage of our approach for multi-document summarization is to consider proper nouns and newly-coined words because we exploit tag clusters of the folksonomy system when detecting key words or analyzing the semantic relatedness among them. Finally, through various experiments on TAC 2008 and TAC 2009 datasets, we demonstrate the superiority of our multi-document summarization technique.

Acknowledgments. This work was supported by the ETRI R&D Program of KCC(Korea Communications Commission), Korea [11921-03001, "Development of Beyond Smart TV Technology"].

References

1. Mani, I.: Automatic Summarization. John Benjamins (2001)
2. Barzilay, R., McKeown, K.R., Elhadad, M.: Information fusion in the context of multi-document summarization. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, ACL 1999, pp. 550–557. Association for Computational Linguistics (1999)
3. Knight, K., Marcu, D.: Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence* 139, 91–107 (2002)
4. McKeown, K.R., Klavans, J.L., Hatzivassiloglou, V., et al.: Towards multidocument summarization by reformulation: Progress and prospects, pp. 453–460. John Wiley & Sons Ltd. (1999)
5. Hennig, L., Labor, D.: Topic-based multi-document summarization with probabilistic latent semantic analysis. In: Proceedings of the International Conference RANLP, pp. 144–149 (2009)
6. Wan, X., Yang, J.: Multi-document summarization using cluster-based link analysis. In: Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, pp. 299–306. ACM (2008)
7. Dang, C., Luo, X.: WordNet-based Document Summarization. In: Proceeding of the 7th WSEAS International Conference on Applied Computer & Applied Computational Science (ACACOS 2008), pp. 383–387 (2008)
8. Zhu, J., Wang, C., He, X., et al.: Tag-oriented document summarization. In: Proceedings of the 18th International Conference on World Wide Web, WWW 2009, pp. 1195–1196. ACM (2009)
9. Kleinberg, J.M.: Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)* 46, 604–632 (1999)
10. Lin, C.Y., Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language, NAACL 2003, pp. 71–78. Association for Computational Linguistics (2003)

ShareDay: A Novel Lifelog Management System for Group Sharing

Lijuan Marissa Zhou¹, Niamh Caprani^{1,2}, Cathal Gurrin¹,
and Noel E. O'Connor²

¹ CLARITY: Centre for Sensor Web Technologies, Dublin City University

² School of Electronic Engineering, Dublin City University

{mzhou,ncaprani,cgurrin}@computing.dcu.ie, oconnorn@eeng.dcu.ie

Abstract. Lifelogging is the automatic capture of daily activities using environmental and wearable sensors such as MobilePhone/SenseCam. The potential to capture such a large data collection presents many challenges, including data analysis, visualisation and motivating users of different ages and technology experience to lifelog. In this paper, we present a new generation of lifelog system to support reminiscence through incorporating event segmentation and group sharing.

Keywords: Multimedia System, Group Sharing, Lifelogging, Touch Screen.

1 Introduction

With the recent availability of wearable sensing technologies and an acceptance of personal data gathering and on-line social sharing (e.g, on Facebook timeline), lifelogging has become a mainstream research topic. We now have the ability to gather and store large volumes of personal data using an inexpensive smart phone. However, with many available lifelogging tools, how to collect, organize and represent lifelog data is still under much discussion [1,2].

Furthermore, people have always collected mementos over lifetime. With the digitization of mementos (photos and videos etc.), researchers have begun to realize the benefit of this to support reminiscence[3]. Sharing digital information is already commonplace, through emails, mobile phones and social networks. However, sharing lifelog data between family members, to our knowledge, has not yet been looked at. Shared reminiscence between family members can serve many functions such as maintaining memories of past relatives, creating bonds and teaching younger family members from the elders' experiences. We believe that sharing lifelogs within a family would enrich reminiscence and story-telling. In this paper we describe a novel software system to support sharing lifelog.

2 ShareDay

A previous study on intergenerational sharing [4] has shown that both older and younger people were more likely to wear a lifelogging device for the purpose of

sharing images rather than simply wearing a lifelog device for private browsing or reminiscence. ShareDay was designed to support browsing and sharing through lifelogs on cross-platforms. To support family reminiscence we have designed the system to be used on a touch screen device displayed in a communal area at home so that all family members can upload, view and share their lifelogs. Users can also view specific person's shared data by clicking on their profile (see Figure 1). The lifelog data can be viewed in two modes: personal and family. In family view, all group members can see all daily events and images in an overall, shared or favourite view. In personal view, only the logged-user can upload data and view their data. If logged, users can manage (share, mark as favourite or delete) and browse their visual lifelogs. There are three main functions incorporated into the ShareDay system: managing personal data, shared data and favourite data. Fig.1 is the main view of the system, which shows all events gathered by all family members in that day. These functions will be described as follows.

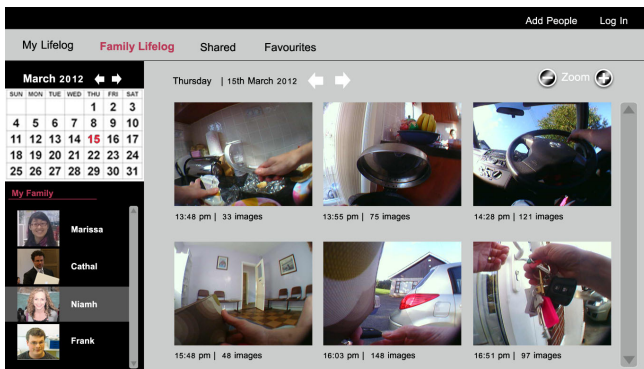


Fig. 1. ShareDay System Overview: Family View

2.1 Managing Personal Lifelog Data

To collect a personal lifelog, the user must initially capture images and upload them to the system. Visual lifelogs present a challenge for developers as they need to represent the users day accurately and in a user-friendly manner, without requiring the user to browse through up to 5,000 images per day. We integrated an event segmentation model[5], which organizes a sequence of SenseCam images into a set of events. Events represent daily activities such as walking, eating, shopping, talking, etc. Keyframe images representing events are selected and displayed for each event, with six large keyframe images being selected for each event. When a user is logged in they can manage (share, favourite or delete) and browse through their visual lifelog. With regard to sharing, logged user can share/mark favourite their daily events/images, so other family members can see shared data, which is designed under the concern of privacy.

2.2 Sharing Lifelog Data

The initial screen of the system displays shared lifelog of each family member. The user can touch on the name of their family member to view lifelogs. A user can also browse through shared lifelogs organized chronologically by touching the *Shared* tab. Group sharing to support family reminiscence is the primary aim of the proposed lifelog management system. However, we also wanted to ensure that users had control over their own lifelogs as the content can be extremely personal. To accommodate for this users can select to share images/events when they are logged into their accounts. These images will automatically be transferred to communal lifelog data set which all members of the family can view.

2.3 Favourite Lifelog Data

In our previous studies the participants reported that when they wanted to share images they had difficulty finding the images due to the vast data set accumulated[2,4]. An easy fix for this was to provide users with a *Favourites* button. These selected images are automatically duplicated into a favourites folder which users can find on the main menu bar.

3 Discussion and Conclusion

We demonstrate a novel lifelogging system for group sharing, which emphasizes on sociability of SenseCam users. In future, we will conduct more user studies on influence of group sharing/favourite to personal lifelog and reminiscence.

Acknowledgement. This research is supported by SFI(11/RFP.1/CMS/3283) and Embark Initiative.

References

1. Caprani, N., Doherty, A.R., Lee, H., Smeaton, A.F., O'Connor, N.E., Gurrin, C.: Designing a touch-screen sensecam browser to support an aging population. In: CHI Extended Abstracts, pp. 4291–4296 (2010)
2. Zhou, L.M., Gurrin, C.: A survey on life logging data capture. In: SenseCam 2012: 3rd Annual Symposium, SenseCam 2012 (2012)
3. Peesapati, S.T., Schwanda, V., Schultz, J., Lepage, M., Yae Jeong, S., Cosley, D.: Pensieve: supporting everyday reminiscence. In: CHI, pp. 2027–2036 (2010)
4. Caprani, N., Gurrin, C., O'Connor, N.E.: Sharing as a motivation for lifelogging. In: SenseCam 2012: 3rd Annual Symposium, SenseCam 2012 (2012)
5. Doherty, A.R., Conaire, C.O., Blighe, M., Smeaton, A.F., O'Connor, N.E.: Combining image descriptors to effectively retrieve events from visual lifelogs. In: Multimedia Information Retrieval, pp. 10–17 (2008)

Helping the Helpers: How Video Retrieval Can Assist Special Interest Groups

Frank Hopfgartner, Jinlin Guo, David Scott, Hongyi Wang, Yang Yang, Zhenxing Zhang, Lijuan Marissa Zhou, and Cathal Gurrin

CLARITY: Centre for Sensor Web Technologies
Glasnevin, Dublin 9, Ireland

{fhopfgartner, jguo, dscott, hwang, yyang, zzhang, mzhou, cgurrin}
@computing.dcu.ie

Abstract. Given the increasing broadcasting data and the ever decreasing spare time that we can spend on consuming this data, systems are required that assist us in identifying important content. Following a use case of a fictional social worker, we introduce a video retrieval system that is designed to assist special interest groups in their information gathering task.

Keywords: video retrieval, demo, broadcasting data.

1 Introduction

“Patrick Murphy is an Irish-American street worker from a small town in the American West. Due to the critical financial situation in his state, the local government is forced to cut the annual budget significantly. His supervisor told him that right now, various boards discuss on which social programs will be shut down forever. Being aware that his clients, mostly poor and homeless people who hardly voice their opinion, will suffer most from these financial cuts, he decides to raise awareness of their needs, thus lobbying for the prosecution of the most important programs. Unfortunately, being on the street every day, he hardly can find time to attend all publicly accessible board meetings. Luckily, the government of his town runs a television channel where they provide an overview over every day activities of the government and their boards.”

In this demo, we showcase our video content organization system that allows fictional Patrick Murphy to easily assess the recordings of these board meetings. In Section 2, we introduce the data corpus that is used. Section 3 describes the required data processing steps. In Section 4, we describe the graphical user interface of our system. Section 5 concludes this demo paper.

2 Data Corpus

Addressing our user scenario, we focus on a city-wide government television channel based in California. Programs include, amongst others, coverage of board

meetings, local press conferences, and commission meetings. Further, information about government services are televised.

3 Data Processing

An important step for easing access to a video corpus is to segment it into semantically coherent segments. In a data corpus that mainly consists of meetings and press conferences, we consider speaker changes to be the semantic segmentation unit and segment the broadcasts accordingly. Further, we detect shot boundaries within these speaker segments and extract the middle key frame to visualize the content of the shot. Similar to Hopfgartner and Jose [1], we extract named entities (persons, locations, organizations) from the recorded closed caption signal, since they provide the highest content load, thus indicating the main subject of the segment. We argue that these entities can be used to filter search results. Moreover, applying subjectivity cues identified by Wilson et al. [2], we determine contextual sentiments on a textual phrase-level. Finally, we use Solr to index all segments, treating the textual transcript, metadata, sentiments, key frame and video URIs and named entities as separate fields in the index.

4 Interaction Using a Graphical User Interface

Figure 1 shows a screenshot of the developed graphical user interface. On the top of the interface, the users can type in a textual search query. Search results, ranked using TF.IDF, are displayed on the right hand side of the interface. Each result is displayed by a representative key frame and a query-biased text snippet. This allows the user of the system to get an initial impression of the content of the retrieved segment. By clicking on one of the results, another window will be opened (not shown on the screenshot) where the video can be played back and neighbored shots are displayed. Further, a color-coded time line is displayed, visualizing passages in the video with negative and positive sentiment, respectively. This display allows the user to assess the general sentimental tone of the video. For example, if our fictitious Patrick Murphy views a board meeting about one of his social projects, the sentiment bar allows him to easily identify those board members who speak in favor and those who speak against it. On the left hand side of the interface, the user can re-define their search query. Given the importance of broadcasting time in the channel, the interface provides facilities to set the broadcasting time. Further, on the bottom left hand side, the users can exclude search results that contain certain named entities. Further down, not visible on the screenshot, is a tag cloud which displays the most frequent entities of the search results. This allows the user to gain a quick overview over the retrieved search results without inspecting them in detail.

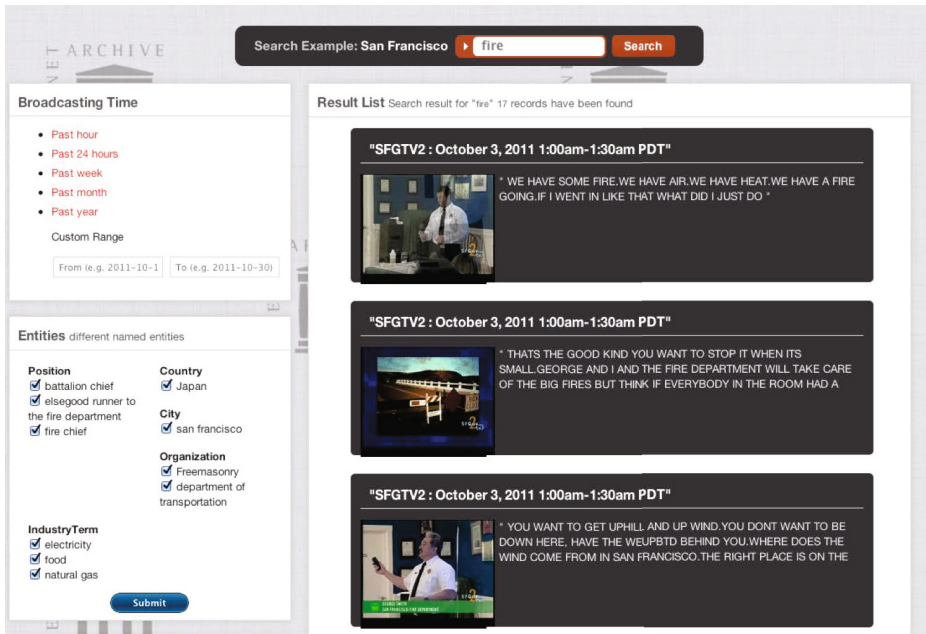


Fig. 1. Screenshot of the Search Interface

5 Conclusion

In this paper, we introduced a video browsing system which is designed to assist users in accessing audio-visual recordings of board meetings. The system allows to access the data corpus of a small video collection. Filtering techniques such as broadcasting time, the appearance of certain named entities and a sentiment analysis ease access to this heavily speech biased data corpus.

Acknowledgments. This research was supported by the Norwegian Research Council (CRI number: 174867) and Science Foundation Ireland under Grant No.: 07/CE/I1147.

References

1. Hopfgartner, F., Jose, J.M.: Semantic user profiling techniques for personalised multimedia recommendation. *Multimedia Syst.* 16(4-5), 255–274 (2010)
2. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: *HLT 2005*, pp. 347–354. *ACL* (2005)

Browsing Linked Video Archives of WWW Video

Zhenxing Zhang, Cathal Gurrin, and Jinlin Guo

CLARITY: Centre for Sensor Web Technologies
Glasnevin, Dublin 9, Ireland

{zzhang, cgurrin}@computing.dcu.ie, jinlin.guo2@mail.dcu.ie

Abstract. In this paper, we describe an interactive video browsing system based on a graph of linked video objects. The system automatically organizes unstructured video archives by exploiting visual content similarity between objects in the videos. By generating a video link graph, the system can conceptually group the videos that contains same objects together for searching and browsing. Both the chosen measures of video object similarity and the video data mining technologies are discussed here and included in the related software demonstrator. In addition, the software offers a query-by-image-example video search capability to jump into the video graph at a certain point to begin browsing the archive.

1 Introduction

Recently, various research efforts have been carried on developing new approaches to object retrieval in large image collections [1]. By applying the efficient text-based query mechanisms, videos containing similar objects can be accurately retrieved from a large dataset in an efficient manner. In this work, we examine new opportunities to study the relationship of videos in a large collection. Instead of using a conventional semantic concept classifier to identify the semantic concepts from videos, our technique links videos based on the presence of objects/feature points within the keyframes of the video content. Automatically linking these videos can support efficient and extensible indexing, fast linkage generation and gives users both links to related content as well as a complete and clear picture of the relationship graph for the entire video collections. Useful information can consequently be summarized together to help users understanding, browsing, and searching of these videos.

2 Indexing

Our aim is to link videos that contain the similar visual entities together, taking into account variances in terms of scale, capture angle, illumination or color appearance. An efficient and accurate indexing and searching algorithm is needed. In this section, the methods and algorithms we employed to address these problems will be discussed in detail.

2.1 Dataset

For this work, a subset of the video collection (about 5,000) for the TRECvid 2012 instance search task was employed, which is composed of user generated videos; hence it is unstructured video data and is likely to have few descriptive annotations. It was originally designed for the task of finding more video segments of a certain specific person, object, or place, given a visual example. This was considered an ideal archive for this prototype video browser for linked archives.

2.2 Video Structure Parsing and Keyframe Selection

In order to automatically organizing a large and unstructured video data collection such as the archive we use here, a shot based segmentation method has been employed to logically divide each video to different shots and one keyframe is selected to represent each shot. This set of keyframes extracted from videos can then be used as still images in large database and will be the subject of the feature extraction and linkage techniques described below.

2.3 Keyframe Representation

By employing a “bag of visual words” model [1], each keyframe is represented as a vector of visual words. For each keyframe in the archive, the affine-invariant Harris-Laplace regions are detected using the technique provided by VLFeat [2]. These regions are the stable areas that are invariant to viewpoint, illumination and scale changes. A 128 dimension SIFT descriptor [3] is then generated based on these regions which will be used in the vector quantization process. We randomly select 20 million descriptors to generate the codebook. The approximate k-means algorithm of [1] has been employed to do the clustering. Each descriptor in keyframes is then assigned with the nearest visual word (cluster center) using the approximate nearest neighbor method. It is important to mention that there will be a quantization error during assignment of descriptors to visual words. Two descriptors that have no similarity could be assigned to same visual words and a link at the keyframe level. This problem will be addressed in detail in next section.

2.4 Link Two Videos with Spatial Verification

Based on the visual words, the link between any keyframe and another will be calculated using L2 distance. Using a conventional TF-IDF weighting scheme, we can tune the performance by reducing the contribution of commonly occurring visual words. For each keyframe, we retrieve the top 1,000 ranked keyframes. Random Sample Consensus (RANSAC) [1] algorithm is applied to solve the quantization error by verifying the spatial consistence between frames. False matching visual words will be filtered out because they don't follow the affine transform rule that the true matches will. To implement RANSAC, we randomly choose 4 match pairs to estimate the affine transformation parameters, and do this for 100 times to get the best model.

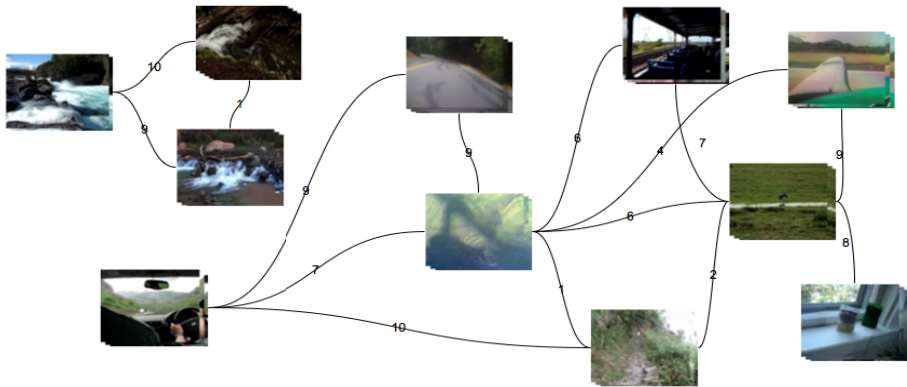


Fig. 1. This figure shows a sample view of the link video graph. Every node in this graph stands for a video from the dataset. The key frames from videos are used for quick view of the video content.

The result of spatial verification can be viewed in the linkage graph presented in Figure 1.

3 The Link Graph and Browser

The video link graph is the index upon which the demonstration software is built. In the video link graph, each node stands for a video in the data collection and weighted edges reflect the (strength of the links) between videos. Multiple links between keyframes in similar videos will result in higher strength links between videos. The browser software is web based to support easy access. To find a good point to begin browsing the archive, the user may submit a photo/image which is processed (as described above for keyframes) to find the most similar videos and this becomes the starting point. When viewing a video, links are provided to the top ranked videos to allow traversal of the linkage graph. Future work involves deploying linkage algorithms (such as a variation of PageRank) over the linkage graph.

References

1. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: Proc. CVPR (2007)
2. Vedaldi, A., Fulkerson, B.: VLFeat - An open and portable library of computer vision algorithms. In: ACM International Conference on Multimedia (2010)
3. Lowe, D.: Distinctive image features from scale invariant keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)

Multi-camera Egocentric Activity Detection for Personal Assistant

Longfei Zhang¹, Yue Gao², Wei Tong³, Gangyi Ding¹, and Alexander Hauptmann³

¹ School of Software, Beijing Institute of Technology, Beijing, China

² School of Computing, National University of Singapore, Singapore

³ School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

{longfeizhang, dgy}@bit.edu.cn, kevin.gao@gmail.com

{weitong, alex}@cs.cmu.edu

Abstract. We demonstrate an egocentric human activity assistant system that has been developed to aid people in doing explicitly encoded motion behavior, such as operating a home infusion pump in sequence. This system is based on a robust multi-camera egocentric human behavior detection approach. This approach detects individual actions in interesting hot regions by spatio-temporal mid-level features, which are built by spatial bag-of-words method in time sliding window. Using a specific infusion pump as a test case, our goal is to detect individual human actions in the operations of a home medical device to see whether the patient is correctly performing the required actions.

Keywords: Multi-camera egocentric, action detection, assistant system.

1 Introduction

In order to maintain aging patients independence, individuals and/or their at-home companions must be able to prepare and apply medical devices accurately and reliably. Since the consequences of making an error in the operating procedure can be literally life-threatening, our system was developed to detect errors and eventually alert patients when patients are preparing or operating medical devices.

Egocentric action analyzing, which aims to employ computers for assisting individuals in everyday life, has become feasible for a broader range of applications due to the increased processing capacity of mobile and wearable devices as well as computers embedded in everyday objects [1,2]. One of typical egocentric capture paradigms is to mount a camera on the head of a subject and record activities from an egocentric perspective (i.e. from the subject's own point of view) [1]. Another is to mount multiple cameras around the person, including ahead, front, side and higher position for more accurate activity understanding results [2]. Our system aims to understand what the user is working on at the present time and what goals the current activities are directed towards. Such as in home medical devices using scenario(as showed in Fig. 1).

There are two key problems to be solved. One is detection accuracy, the other is that the system needs to alert in time when wrong operation happens. For the first problem, a flexible mechanism is required so that the system could learn and refine representations of high-level tasks from observation and interaction with a human operator, based on a set of underlying primitive actions that the system already understands [3]. In this

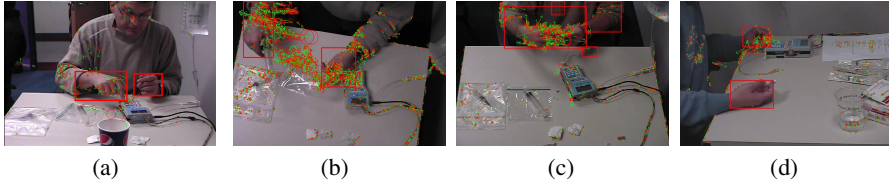


Fig. 1. Examples of egocentric actions with MoSIFT feature points from infusion pump dataset: (a) is from the “Front” camera, (b) is from the “Veryhigh” camera, (c) is from the “Above” camera and (d) is from the “Side” camera. Red rectangles are the hot regions. Arrows in green and red circles are MoSIFT points.

case, spatial selection of attention [4] can help us to locate these actions in spatial when we build spatial bag-of-words event detecting model. For the other problem, a well designed GPU enhanced programming framework is needed.

The technical work in this paper consists of two components: (1) training the multi-camera egocentric action model by recording a set of operations, (2) observing a new instance of the operation actions and recognizing what operation is being performed.

2 Egocentric Action Detection

Algorithm 1 gives the details of our approach.

Algorithm 1. Multi-camera egocentric action assistant model.

The input is a egocentric video sequence $V = (V_0, V_1 \dots V_k)$. The output is a respond of operation $R(Y, N)$. Hands are initialized to a color and HOG histogram H . Define $M = (M_0, M_1 \dots M_m)$ as category of action. Back projection is performed to find the hot region in egocentric video.

Initialize $q=1, H$;

repeat

- Step1. construct hot region a_0^q by back projection algorithm with histogram H , a_0^q belongs to A , $A = (a_0^q, a_1^q \dots a_k^q)$.
- Step2. extract MoSIFT points p_0^q inside a_0^q .
- Step3. construct bag-of-words b_0^r inside temporal sliding window t_0^r , $r = 1..k - 1$.
- Step4. classify the b_0^r as a M_i .
- Step5. get $R(Y, N)_k$ if M_i is not in right operating sequece.

until $q=k$;

Step6. Fuse $R(Y, N)_k$ to $R(Y, N)_k$ to get the finial decision $R(Y, N)$

Return $R(Y, N)$

In our approach, we convert the video from a volume of pixels to compact but descriptive interesting points. We employ Chen’s MoSIFT detector [5] to detect and describe spatio-temporal interesting points.

For each key frame, the number of extracted key points can be different. A spatial bag-of-words (BoW) approach, which is based on a selective sampling k-means clustering and hot region detecting, is used to quantize the combination of motion and appearance features to a fixed length vector for each frame. A χ^2 kernel SVM classifier

is applied because it has been shown to be better for calculating histogram distances and directly compare to previous work [2].

We briefly introduce the MoSIFT [5] as follows. The MoSIFT feature integrates SIFT and optical flow in multiple scales to represent spatio-temporal information. MoSIFT feature point, which is generated by combining SIFT RANSAC matching and optical flow computing, is more robust to describe the motion information than traditional optical flow.

3 Experiments

The core ability of our system is to monitor patients egocentric operations and eventually alert them when errors happen. We use infusion pump operation as a test case and the Pump dataset as training dataset.

In the Pump dataset [2], there are 6 subjects. When operating the infusion pump, subjects are required to follow a protocol with 22 operations. Four cameras record activities from different views (as showed in Fig. 1): front, side, overhead and diagonally above the head, which is labeled as “Veryhigh”. Classification results from 4 cameras are fused and finally the averaged recognition rate is over 60% with average frame rate over 10 FPS (Frame per second).

4 Future Plans

We are currently working on improving the interesting region detection especially action objects detection automatically and optimizing action patterns. Specifically to the focus of this demonstration, we are exploring an approach based on learning by demonstration that allows the user to perform the activities without initialization. Ongoing work seeks to improve the robustness, reduce the respond times of the action detection and make the user interface friendly.

Acknowledgments. This material is based in part upon work supported by the National Science Foundation Grant IIS-0917072, and by the National Institutes of Health Grant 1RC1MH090021-0110. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the National Institutes of Health.

References

1. Fathi, A., Ren, X., Rehg, J.: Learning to Recognize Objects in Egocentric Activities. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR (2011)
2. Gao, Z., Detyniecki, M., Chen, M.-Y., Hauptmann, A.G., Wactlar, H.D., Cai, A.: The Application of Spatio-temporal Feature and Multi-Sensor in Home Medical Devices. *International Journal of Digital Content Technology and its Applications(IJDCTA)* 4(6), 69–78 (2010)
3. Laptev, I.: On space-time interest points. *International Journal of Computer Vision* 64(2-3), 107–123 (2005)
4. Fan, J., Wu, Y., Dai, S.: Discriminative Spatial Attention for Robust Tracking. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 480–493. Springer, Heidelberg (2010)
5. Chen, M.Y., Hauptmann, A.: MoSIFT: Recognizing human actions in surveillance videos. CMU-CS-09-161, Carnegie Mellon University (2009)

Music Search Engine with Virtual Musical Instruments Playing Interface

Mei Wang¹, Wei Mao², and Hai-Kiat Goh³

¹ Donghua University, Shanghai, China

² ZTE Corporation, Shanghai, China

³ KAI Square Pte Ltd, Singapore

Abstract. In this paper, we presents a novel music search engine with query by playing the virtual musical instruments. Different from the previous query-by-keywords or query-by-hamming methods, the proposed search engine provides a new input interface, which allows the user to play simulated musical instruments to obtain the audio clip to do search. Since the sounds by playing certain musical instrument have the common standard, query-by-playing can effectively reduce the gap of users' intention and input signals. In the other hand, search in this way can provide more possibilities for different kinds of people especially for the professionals to accurately retrieve more different types of music.

1 Introduction

With the rapid development of the audio industry, music information retrieval (MIR) has been an active research area in multimedia communities in recent years. Many previous works have been done on MIR system for content-based music searching, however the problems are still challenging.

Among most existing music information retrieval systems, multiple novel input modalities have been used to fill users intention gap [1,5]. For example, in query-by-example application [3], users can record a piece of song as query and the system then retrieves the identical songs from the server. In query-by-humming system [4], a user can sing into the microphone and the system can search for a song that is similar to the user's humming. Query-by-tapping [2] system allows a user to tap the mouse or clap into microphone to search based on the rhythmic pattern. In these methods, query by humming/sing favorably fits the practical need, especially when the user forgets the title and the singer of the song they would like to find them in a large music database. It probably was one of the core techniques in content-based music information retrieval. There are always two steps to search similar music after the singing or humming audio signals obtained. The first one is to extract an abstract description features of the audio signal which reflects the perceptually relevant aspects of the signals, followed by the step to match the extracted information based on the distance function. However, considering the casual and unprofessional singing process, some uncertain factors and noise will be naturally introduced into the user' input, such as the inconsistent tune, inaccurate tone, insufficient duration



Fig. 1. The input interface of the proposed music search engine

and so on, which will greatly harm the accuracy of feature extraction and feature matching, leading to the inaccurate retrieval results.

In recent, touch screen technology has been widely used in the popular digital equipments such as tablet devices, IPAD or large-screen telephones. The touch screen technology provides the practicality for the user to simulate playing different kinds of musical instruments on these equipment smoothly. At the same time, the MIDI standards and protocols give support to receive, analyze and output the simulated playing signals.

Based on these technologies, in this paper, we propose a novel music search engine with query by playing the virtual musical instruments. Different from the previous work, which input the keywords such as song name, singer name, or hamming signals to search music, the proposed search engine provides a new interface, which allows the user to play different virtual music instruments to achieve the music clip to do search. Since the sound of the same instrument have the common standard, query-by-playing can effectively reduce the noises introduced in the casual and unprofessional singing/hamming process, correspondingly reducing the gap of users intention and input signals and improving the retrieval performance. Although the proposed interface increase some professional demand for the user. However, in the other hand, search in this way can provide more possibilities for different kinds of people especially for the the people with musical foundation to accurately retrieve more different types of music. For example, some people with music foundation can search classical music, which is not well handled in the previous search engines.

2 System Implementation

The proposed search engine mainly consists of three parts: input interface, feature extraction and feature matching. Fig. 1 illustrates our search input interface. There are three panels in the interface. The upper left is the instrument selection panel. The upper right is the command button panel. The center is the instrument playing region. An example of the simulated piano is provide in the figure. The user first selects the favorite instrument, and then clicks the “play” button. In the instrument playing region, the music clip will be played and input. Finally, after clicking the “Search” button, the search results ranking with the

feature matching distance will be returned to the user. The detailed information will be introduced in the following.

Input Interface: we provide some simulated musical instruments that are easy to be played with touch screen, for example piano, gita and guzheng. We implement the virtual instruments based on MIDI technology. Taking piano as an example, we bind the simulated piano key to the MIDI channels. When the user playing, the notes will be sent to the corresponding channels, and then we can obtain the music clip with MIDI format.

Feature Extraction: in this part, we extract the abstract description features of the audio signals. We parse input MIDI file and then extract the pitch, melody information, and then obtain the feature vector.

Feature Matching: here we use DTW [6] algorithm for distance matching. When matching with the MIDI file in the music collections, we extract the same type of features with the same method. While matching with other formats (mp3, wma and so on), we can convert other types to MIDI format first or extract the features directly from the music files. Query-by-playing can effectively reduce the noise induced by the hamming process, so the matching accuracy will be improved significantly.

3 Conclusion

In this paper, we proposed a music search engine with query by playing the virtual musical instruments. Different from the previous work, the proposed search engine provides a new input interface, which allows the user to play different simulated musical instruments to obtain the audio clip to do search.

Acknowledgments. This work was partially supported by the Natural Science Foundation of China Grant No.61103046, the Natural Science Foundation of Shanghai Grant No.11ZR1401200.

References

1. Typke, R., Wiering, F., Veltkamp, R.: A survey of music information retrieval systems. In: ISMIR, pp. 153–160 (2005)
2. Hanna, P., Robine, M.: Query by tapping system based on alignment algorithm. In: ICASSP 2009, pp. 1881–1884 (2009)
3. Wang, A.: The Shazam music recognition service. *Commun. ACM* 49(8), 44–48 (2006)
4. The midomi music search, <http://www.midomi.com>
5. Wang, M., Hua, X.S.: Active Learning in Multimedia Annotation and Retrieval: A Survey. *ACM Transactions on Intelligent Systems and Technology* 2(2), 10–31 (2011)
6. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing* 26(1), 43–49 (1978)

Navilog: A Museum Guide and Location Logging System Based on Image Recognition

Soichiro Kawamura, Tomoko Ohtani, and Kiyoharu Aizawa

The University of Tokyo, 7-3-1 Hongo Bunkyo Tokyo 113-8656, Japan
(kawamura,fritz.tmk,aizawa)@hal.t.u-tokyo.ac.jp

Abstract. We developed a computer vision-based mobile museum guide system named “Navilog”. It is a multimedia application for tablet devices. Using Navilog, visitors can take a picture of exhibits, and it identifies the exhibit and it shows additional descriptions and content related to it. It also enables them to log their locations within the museum. We made an experiment in the Railway Museum in Saitama, Japan.

1 Introduction

Recently, information technology for museum navigation is advancing: Wi-Fi based localization is utilized in some museums, and the visitors can easily choose the exhibits using their smartphones, and read the additional information. However, the museum needs to implement additional equipment that is many Wi-Fi stations, for the localization.

We made a system named “Navilog” which works on small tablet devices. The functions of system are identifying exhibits by image processing and showing additional information for users. Navilog performs those tasks by analyzing photos of exhibits taken by users.

2 A Museum Guide Prototype

In our system, the tablet device performs all image processing. The image database of museum exhibits is not large, so it can be stored in storage of the mobile device. This enables to make search executable without network connection. Such standalone implementation makes operation easy. We can use existing high performance descriptors without addition of data compression.

We made a prototype (Fig. 1) and asked museum visitors to evaluate its usability. Guide device runs along following steps.

1. Take query photograph of the museum exhibit (by the user)
2. Select region of interest (by the user)
3. Detect interest points in query
4. Extract descriptors
5. Match feature of query with those in database

First, the user takes photo of exhibition by camera in a tablet. Photos in database we took when the museum is closed, so these photos contain only exhibit without visitors. However, the photos taken by users are likely to include other visitors. Then, the query photo might not match with those in the database. To resolve this problem, the user trims the exhibit region of the photo. After photo is taken, the device shows its picture on the display and the user interactively selects the region of interest. The user marks the exhibit using a touch operation (Fig. 2) and the system determines its bounding box of this region.

The system searches for the query image in the database. The pictures in the database have been taken from various positions and angles for each exhibit.

First, the system detects interest points in the query image. We use oFAST detector in ORB[1]. It is multi-scale fast detection algorithm.

Next, it extracts the feature descriptor around the interested points. We use SURF[2] which uses box filter and integral image, and is faster than SIFT. SURF describes image patch as 64 dimensional vectors. One query image has some hundred interested points. So it is computationally intensive to match descriptors of the query with these from database.

We use Bag-of-Features[3] method. We made “Visual word” vectors from images in the database. The pictures in database are converted to visual words histograms. These histograms are stored in the storage in the tablet device.

When the query image is shot, the system calculates its BoF histogram and compares those in database. Feature vectors common in most images are useless for image categorization so these should have less emphasis. *tf-idf* is a weighting method based on frequency.

When histograms are compared, these are weighted by *tf-idf* and then compared by L1 norm. System sorts these results in order of distance, allows the user to select proper one (Fig. 3).

When a user selects an exhibit from the list of results, Navilog shows the corresponding contents (e.g. its title, factory and guide movie).

Taken photos are showed in thumbnail view. The viewer displays photos and brief captions of exhibits. In order to read detailed explanation, the user taps button on the bottom of photo (Fig. 4).



Fig. 1. Museum guide prototype “Navilog”



Fig. 2. Selection of region of interest



Fig. 3. Result of image search

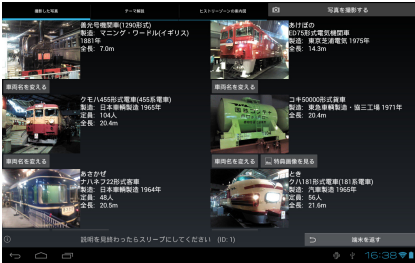


Fig. 4. User's photos and descriptions

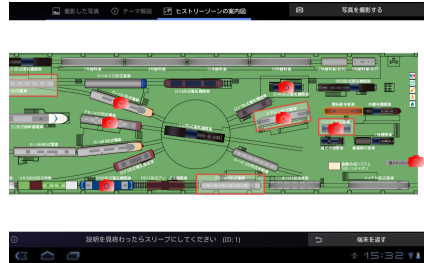


Fig. 5. Camera icons show visited

The database also has coordinates of exhibition in exhibition room. The system shows the user's location history on the museum map (Fig. 5).

3 Conclusion

We developed a new museum guide using image recognition and conducted a user study to evaluate it.

In this research, the system determines foreground and background by touch-screen operation and reduces computational cost as all operations are done in mobile device. It requires neither networks nor servers.

We intend to improve this system for smaller devices such as smartphones.

References

1. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: Orb: An efficient alternative to sift or surf. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2564–2571 (2011)
2. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). *Comput. Vis. Image Underst.* 110(3), 346–359 (2008)
3. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: *Proceedings of the Ninth IEEE International Conference on Computer Vision*, vol. 2, pp. 1470–1477 (2003)

Early Skip Mode Detection by Exploring Extra Skip Patterns for H.264 Coarse Grain Quality Scalable Video Coding

Hao Zhang, Xiao Yu Zhu, and Xuan He

School of Information Science and Engineering
Central South University, Changsha, Hunan 410083 China
hao@csu.edu.cn

Abstract. We propose a fast early skip mode detection approach by exploiting extra skip patterns that are missed by previous research. With a larger early skip candidate set, a larger number of macroblocks at the enhancement layer could be detected as skipped mode. Experimental results demonstrate that, the proposed method achieves higher encoding time reduction with negligible quality loss compared with previous research results.

1 Introduction

SVC requires coding of both base layer (BL) and enhancement layer (EL) compared with H.264 single layer video coding. In the literature, various fast mode decision (FMD) algorithms have been proposed to reduce H.264 computational complexity. Recently, Shen *et al.* focused on efficient SKIP mode detection for CGS as this mode is very popular in SVC EL [1]. Their approach decided a MB at EL to be a SKIP MB if the co-located MB at BL, the top MB and the left MB are all encoded as SKIP mode. We find there is still much room for improvement in SKIP mode detection for CGS. In this paper, we propose a fast early SKIP mode detection by considering other skip patterns, RD costs and CBP values.

2 Proposed Early SKIP Algorithm

we define SKIP RD cost as the RD cost resulting from only checking SKIP mode. $C(x)$ is used to represent the SKIP RD cost for any MB x . Please note that $C(x)$ is not the RD cost of the best mode (minimum RD cost) after the mode decision procedure is finished. It is equal to the minimum RD cost only when SKIP mode has been chosen as the best mode. We define a 3-tuple $V_s = (\text{bskip}(B_c), \text{bskip}(E_l), \text{bskip}(E_u))$, where $\text{bskip}(B_c)$ is a binary variable that is set to one if B_c is a skipped MB and zero otherwise. Our observation could be conveyed by the following equations:

$$C(E_c) \leq \frac{\alpha}{2}(C(E_l) + C(E_u)) \quad (1)$$

$$|C(E_c) - C(E_l)| \leq \alpha |C(E_c) - C(E_u)| \quad (2)$$

$$|C(E_c) - C(E_u)| \leq \alpha |C(E_c) - C(E_l)| \quad (3)$$

In these equations, α is a configurable parameter and offers performance trade-offs. Based on the analysis of many experimental results, we found α can be set to 1.5 to balance the encoding speed and video quality. Eq.1 is for the case of $V_s = (1, 1, 1)$, i.e., all the B_c , E_l , E_u are skipped. In this case, if E_c is skipped, its SKIP RD cost is usually smaller than the average SKIP RD cost of the top and left MBs. Eq.2 and Eq.3 correspond to $V_s = (1, 1, 0)$ and $V_s = (1, 1, 0)$, respectively. The whole procedure is summarized as follows. We use a flag SKIP_FLAG to denote whether this MB can be early skipped or not.

- Step 1) For each MB, if $V_s = (1, 1, 1)$ or $(1, 1, 0)$ or $(1, 0, 1)$, initialize SKIP_FLAG to *true* and goto STEP 2; otherwise goto to STEP 10 for the normal mode decision process.
- Step 2) Calculate the RD cost of SKIP mode. Note that for B frames, Direct16x16 mode is performed first and the SKIP mode could be deduced if the resulted CBP is zero.
- Step 3) If $V_s = (1, 1, 1)$ and the current slice is B-slice, goto STEP 4; otherwise, goto STEP 5.
- Step 4) If CBP is non zero, goto STEP 10; otherwise, goto to STEP 8.
- Step 5) If $V_s = (1, 1, 1)$ and the current slice is P-slice, goto STEP 6; otherwise goto STEP 7.
- Step 6) Set SKIP_FLAG to *false* if Eq.1 is not satisfied. Goto STEP 8.
- Step 7) If $V_s = (1, 1, 0)$, set SKIP_FLAG to *false* if Eq.2 is not satisfied. If $V_s = (1, 0, 1)$, set SKIP_FLAG to *false* if Eq.3 is not satisfied.
- Step 8) If SKIP_FLAG is *false*, goto STEP 10.
- Step 9) Check BL_SKIP mode, and choose the better mode between SKIP and BL_SKIP. Goto STEP 11.
- Step 10) Conduct the normal exhaustive mode decision for the rest of modes and determine the best mode.
- Step 11) Go to STEP 1 to process the next MB.

3 Experimental Results

JSVM 9.19 is used for the two layer CGS encoding. The simulation runs on a PC of 2.2 GHz CPU and 8 GB memory. We use BDBR and BDPSNR to measure the average quality change [2]. Tab.1 lists the results when comparing the proposed approach with JSVM and shen's algorithm. It demonstrates that in average, the proposed approach achieves 32.81% reduction of encoding time with 0.011dB BDPSNR loss comparing with JSVM. For slow motion sequences like "Akiyo", the reduction is about 55% for all DQP settings. Comparing with Shen's algorithm, the proposed approach achieves 11.42% time reduction, 0.035 dB BDPSNR gain with 0.684% BDRATE decrement on average (Tab. 1). Specifically, the maximum time reduction 22.44% is achieved for sequence "Salesman" at DQP = 10 with 0 dB BDPSNR gain and 0.473% BDRATE decrement. Overall, in many cases, the time reduction is more than 10% with negligible quality changes.

Table 1. Comparing the proposed early skip mode decision with JSVM and shen’s algorithm

		JSVM			shen’s algorithm		
Sequence	DQP	BDPSNR	BDBR	ATS	BDPSNR	BDBR	ATS
Mobile	2	0.002	-0.011	19.01	0.002	-0.047	9.83
	5	-0.004	0.173	17.13	0.010	-0.145	7.91
	10	-0.018	0.364	11.28	0.010	-0.166	6.29
Coastguard	2	0.000	0.030	30.14	0.016	-0.495	14.54
	5	-0.009	0.204	27.76	0.029	-0.748	14.08
	10	-0.033	0.745	20.65	0.019	-0.493	11.40
Salesman	2	0.007	-0.076	43.23	0.019	-0.075	21.12
	5	-0.015	0.205	42.62	0.009	-0.009	21.03
	10	-0.031	0.510	41.31	0.000	-0.473	22.44
Akiyo	2	0.024	-0.246	56.52	0.036	-0.400	15.74
	5	-0.028	0.582	56.37	0.025	-0.557	15.51
	10	-0.069	1.805	54.40	0.039	-1.022	15.10
Paris	2	0.009	-0.009	36.49	0.084	-1.354	15.47
	5	-0.014	0.213	35.33	0.083	-1.278	15.23
	10	-0.036	0.641	31.39	0.057	-0.946	13.83
Silent	2	0.014	-0.154	48.43	0.016	-0.206	14.94
	5	-0.010	0.196	48.37	0.018	-0.306	14.70
	10	-0.056	1.062	44.92	0.015	-0.274	13.58
Vidyol	2	0.103	0.003	51.85	0.061	-0.893	9.14
	5	-0.029	0.923	49.85	0.060	-1.652	9.41
	10	-0.058	2.518	41.48	0.054	-2.871	9.89
Shields	2	0.001	0.029	16.05	0.007	-0.250	2.52
	5	-0.004	0.216	13.80	0.045	-1.449	1.95
	10	-0.025	1.455	7.60	0.120	-1.495	-1.11
mobcal	2	0.011	-0.133	16.58	0.010	-0.303	5.52
	5	-0.001	0.135	13.37	0.018	-0.672	4.81
	10	-0.025	1.370	9.97	0.064	0.105	3.43
Average		-0.011	0.472	32.81	0.035	-0.684	11.42

4 Conclusion

We propose a novel early skip algorithm in this paper. In the future, we are interested in exploring other mechanisms to speed up SVC encoders such as fast motion estimation.

References

1. Shen, L., Sun, Y., Liu, Z., Zhang, Z.: Efficient skip mode detection for coarse grain quality scalable video coding. *IEEE Signal Processing Letters* 17, 887–890 (2010)
2. Bjontegaard, G.: Calculation of average psnr differences between rdcurves. VCEG-M33. ITU-T Q6/SG16 (April 2001)

A Video Communication System Based on Spatial Rewriting and ROI Rewriting

Hongtao Wang, Fangdong Chen, Bin Li, Dong Zhang, and Houqiang Li

Information Processing Center, Dept. of EEIS
USTC
Hefei, Anhui, P.R. China
wanght99@mail.ustc.edu.cn

Abstract. Scalable Video Coding (SVC), as an extension of H.264/AVC, has been designed to provide H.264/AVC compatible base layer and spatial, temporal and quality enhancement layers. Bit-stream rewriting in SVC standard makes it possible to convert a quality enhancement layer to a H.264/AVC bit-stream. So that H.264/AVC decoder users could also experience high quality video content when network condition and hardware permits. In this paper, we present a scalable video rewriting system which is featured by the ability to rewrite spatial enhancement layers and range-of-interest (ROI) of enhancement layers. Compared to traditional rewriting, the proposed system is suitable for more application scenarios and is more flexible.

Keywords: SVC, spatial rewriting, ROI.

1 Introduction

H.264/AVC has been widely used and achieved great success since it came out in 2003 [1]. H.264/SVC, which provides coded video stream with spatial, temporal and quality scalability, was approved in 2007 as an extension of H.264/AVC [2]. The emerging of H.264/SVC offers the possibility that one bit-stream could be adaptively transmitted and decoded according to network condition and end user's demand. In a coded SVC bit-stream, only base layer is compatible with H.264/AVC. This means that users with H.264/AVC decoders could only get base layer content even if their bandwidth and processing power allows higher quality of service. On the other hand, when it comes to a specific user, hardware ability is determined and network condition is often relatively stable. So there's no need to transmit a scalable stream to such users. These two problems are often solved by transcoding or rewriting SVC bit-stream to H.264/AVC bit-stream at last routers or media gateways. Rewriting is a tool provided by H.264/SVC to convert a SVC bit-stream into AVC bit-stream. Compared to transcoding, it provides lossless conversion with lower complexity. However, rewriting technique provided by SVC could only be operated on quality enhancement layers. Authors of [3] proposed a hybrid bit-stream rewriting approach to support both spatial and quality rewriting. Region-of-interest (ROI), which allows encoders to encode a specified region of a picture (ROI) as a separate slice, is another tool

supported by H.264/SVC. This tool makes it possible for a SVC decoder user to request only part of an enhancement layer and get a high quality ROI when receiving the whole enhancement layer is not allowed. However, traditional rewriting doesn't support rewriting of ROI, which means that H.264/AVC decoder users could not benefit the ROI technique provided by SVC. In this paper, we present a rewriting system which is able to rewrite ROI of an enhancement layer into a H.264/AVC bit-stream. When an AVC user who could not receive the whole enhancement layer is more interested in a high resolution version of ROI than the base layer content, he could request the ROI of enhancement layer other than the base layer to get a higher quality of service.

2 Spatial Rewriting and ROI Rewriting

In the proposed rewriting system, spatial rewriting and ROI rewriting are two key features. To support spatial rewriting, the scheme proposed in [3] is adopted, which was developed base on the principle of residue up-sampling in transform domain. The computational complexity of this approach is much lower than that of transcoding while RD performance is acceptable.

We also developed a ROI rewriting scheme. In addition to spatial rewriting, a new rewriting technique is proposed to extract ROI from enhancement layer and encapsulate it into a single H.264/AVC bit-stream. Several constraints were applied to the SVC encoder so that samples in ROI could be reconstructed without any drift even if the information of background is totally lost in rewriting process.

3 System Description

As shown in Figure 1, the framework of our demo system is almost the same with that of a typical SVC bit-stream transmitting system, except that the rewriting part is specifically designed to provide spatial rewriting and ROI rewriting as shown in section 2.

First, SVC bit-stream is generated and stored on the video server. In our demo, the bit-stream contains two layers with different spatial resolution. Both width and height of the enhancement layer is twice as large as that of the base layer. In the enhancement layer, a ROI with the same size as the base layer is coded. Positions of ROI are the same through all pictures.

Then, for H.264/AVC decoder users, the SVC bit-stream is transmitted to media gateway, on which correspondent H.264/AVC bit-stream is generated according to their demand through rewriting process. Users are offered the choices of receiving the base layer, enhancement layer and ROI of enhancement layer. If bandwidth and hardware permits, rewritten enhancement layer would be sent to users. Otherwise, clients could choose between base layer and ROI of enhancement layer according to their preference and video content, thus the overall quality of service would be improved.

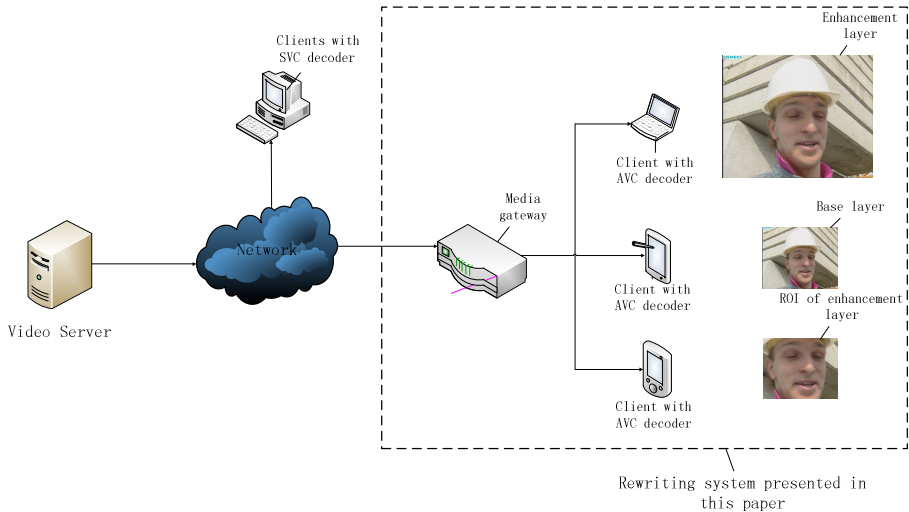


Fig. 1. Structure of rewriting system presented in this paper

4 Conclusions

In this paper, a novel video bit-stream rewriting system has been presented. By supporting spatial rewriting and ROI rewriting, the proposed rewriting system overcomes traditional rewriting technique in both practicability and flexibility.

References

1. 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)
2. Schwarz, H., Marpe, D., Wiegand, T.: Overview of the scalable video coding extension of the H.264/AVC standard. *IEEE Trans. on Circuits and Systems for Video Technology* 17(9), 1103–1120 (2007)
3. Li, B., Guo, Y., Li, H., Chen, C.W.: Hybrid bit-stream rewriting from scalable video coding to H.264/AVC. In: *Visual Communications and Image Processing 2010, Proceedings of the SPIE*, vol. 7744, pp. 77441A–77441A-10 (2010)

NExT-Live: A Live Observatory on Social Media

Huanbo Luan^{1,2}, Dejun Hou¹, and Tat-Seng Chua¹

¹ School of Computing, National University of Singapore, Singapore

² Department of Computer Science and Technology, Tsinghua University, China
luanhuanbo@gmail.com, {houdj, chuats}@comp.nus.edu.sg

Abstract. This demonstration presents a live observatory system named ‘*NExT-Live*’. It aims to analyze live online social media data to mine social phenomena, senses, influences and geographic trends dynamically. It builds an efficient and robust set of crawlers to continually crawl online social interactions on various social networking sites, covering contents from different facets and in different medium types. It then performs analysis to fuse these social media data to generate analytics at different levels. In particular, it researches into high-level analytics to mine senses of different target entities, including People Sense, Location Sense, Topic Sense and Organization Sense. *NExT-Live* provides a live observatory platform that enables people to know the happenings of the place in order to lead better life.

Keywords: Live, Observatory, Monitoring, Social Media, UGC, NExT.

1 Introduction

We are living in the midst of a rich social media environment. We freely and spontaneously generate contents as part of our daily activities including making comments, sharing photos, checking-in to locations, asking and answering questions. Through the wide variety of social media channels, more and more such real-time social media data, collectively known as User-Generated Content (UGC), are being generated. The contents of UGC reflect the pulse of a society and the tone of public opinion, and affect our culture and the way we communicate. Aiming to better understand and analyze live social interactions[1], social media observation and long-term digital preservation have become highly relevant and urgent. Thus there is a strong need for live data crawling, archiving, access, retrieve and analysis.

Web science research community has recently proposed the creation of a global “Web Observatory Community Group” to establish a global open data resource collaboratively by many web observatory nodes across the world [2]. On the other hand, there are some commercial social media monitoring tools and platforms that claim to be able to help track and monitor business or brand in social media channels such as Radian6, BuzzLogic, Visible Technologies, Brandwatch, Brandtology. Although some such tools show good marketing performance, they usually suffer from the problems of narrow application domain, limited data coverage and data types, in which most focus primarily on twitter data. Moreover, most such tools are not fully automated and cannot handle live data well.

To address the above problems, we propose a livesocial observatory system named ‘NExT-Live’ to mine multiple social channels automatically. It can continually crawl live and semi-structured multimedia UGC data, including text, images, videos and user-relation graphs etc. It supports real-time analysis and fusion of these data sources to generate multiple social analytics, including People Sense, Location Sense, Topic Sense and Organization Sense.

2 System Architecture and Implementation

The overall system framework of NExT-Live is illustrated in Fig.1. The system comprises three layers: *Live Data Crawlers*, *Big Data Management* and *Multi-phase Observatory*. NExT-Live currently runs on a cluster with 17 server nodes within the NUS campus.

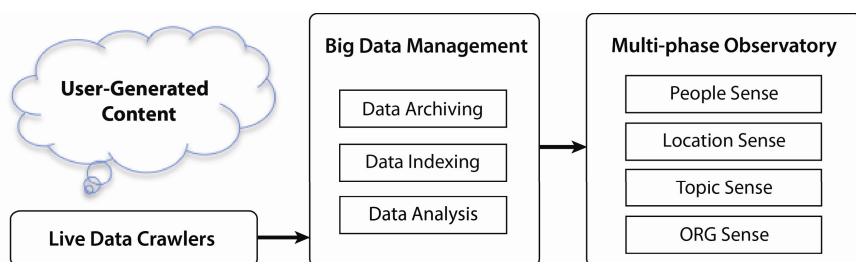


Fig. 1. The overall system architecture of NExT-Live

2.1 Live Data Crawler

NExT-Live tracks multiple social networking sites including *Flickr*, *Foursquare*, *Instagram*, *Panoramio*, *TecentWeibo*, *SinaWeibo*, *Twitter*, *Youtube*, *Amazon*, *Dianping*, *Fantong*, as well as some forum and blog sites. It provides the best real-time coverage of multi-modality UGC such as text posts, user comments, images, videos, user profiles and user relations. In order to ensure continual real-time crawling, we build a set of live robust crawlers that works well across different platforms, channels, and is easy to maintain and extend. The crawlers are made intelligent and robust by supporting IP proxy, heuristically crawling, noise filtering, exception handling, as well as multiple threads and distributed crawling. Table 1 presents a glimpse of the size and variety of live UGC data that we have crawled over a 4½ month period.

Table 1. The crawled User-Generated Content and their sizes (1 May 2012 - 15 Sep 2012)

Data Types	Number of Posts	Size
Micro-blog Posts	1,402,948,496	988 GB
User Comments	175,732,324	178 GB
User Profiles	132,715,138	129 GB
Images	242,913,348	21 TB
Videos	92,277	3.2 TB

2.2 Big Data Management

The live data stream is sent to *Big Data Management* module to perform:

Data Archiving: It utilizes MongoDB and NFS to store text and media data in distributed servers. MongoDB stores JSON-like documents with dynamic schemas and shows good scalability and agility in handling huge data set.

Data Indexing: It then triggers indexing function automatically to build the distributed index in real time for data access and retrieval. The text index is created with SOLR and visual index is generated with hashing and inverted files based on the extracted visual features.

Data Analysis: It carries out the analysis and fusion of multiple UGC sources to generate higher order analytics.

2.3 Multi-phase Observatory

NExT-Live offers multi-phase observatory that helps users better understand the trends and pulses of a society. It builds tools to perform content analysis, data fusion, topic mining, user community discovery, sentiment analysis, as well as the integration of multiple social signals to track and mine events and senses in society. In particular, given a target topic, it mines the evolution of relevant contents, user community and events and integrates them to infer the sense of the entity. The entity can be a person, location, topic or an organization, thus giving rise to observatory for people, location, topic and organization senses. For example, the “Organization Sense Observatory” will analyze relevant UGCs to uncover both emerging and hot events, as well as user community and key users, related to the target organization; while the “People Sense Observatory” will return and analyze what other people post and say about the target person. Collectively, it provides valuable observatories to help us better understand ourselves and the larger environment that we live in.

Acknowledgement. *NExT-Live* system is developed by NExT Search Center [3], which is supported by the Singapore National Research Foundation & Interactive Digital Media R&D Program Office, MDA under research grant (WBS:R-252-300-001-490).

References

1. Cui, P., Wang, F., Liu, S.W., Ou, M.D., Yang, S.Q.: Who Should Share What? Item-level Social Influence Prediction for Users and Posts Ranking. In: International ACM SIGIR Conference (2011)
2. <http://www.w3.org/community/webobservatory/>
3. Chua, T.S., Luan, H.B., Sun, M.S., Yang, S.Q.: NExT: NUS-Tsinghua Center for Extreme Search of User-Generated Content. *IEEE Multimedia* 19(3), 81–87 (2012)

Online Boosting Tracking with Fragmented Model

Dingcheng Shen^{1,2}, Hua Zhang^{1,2}, Yanbing Xue^{1,2,*}, Guangping Xu^{1,2}, and Zan Gao^{1,2}

¹ Key Laboratory of Computer Vision and System, Ministry of Education

² Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology,
Tianjin University of Technology, 300384, Tianjin, China
yanbingxue@163.com

Abstract. We propose a novel method combining online boosting and fragment to overcome the drifting problem in on-line boosting tracking. We find that in previous on-line boosting method, the voting weights of the first few selectors are so big that the remainders can not affect the final strong classifier. This problem occurs because the voting weight of selectors are passing globally to adapt to the object variation, but usually only parts of object changes significantly in short time, and the changing part only affect its neighborhood, not the whole target area. So we divide the selector into fragments to get spatial information. The best weak classifier in each selector is combined linearly to get the final strong classifier and then find the location of the object in next frame. Experiments show robustness and generality of the proposed method.

Keywords: on-line boosting, fragment, voting weight, drift.

1 Introduction

Research in tracking plays a key role in understanding motion and structure of objects. It finds numerous applications including surveillance, human-computer interaction, traffic pattern analysis, recognition, medical image processing. It is a great challenge to design robust visual tracking methods which can cope with the inevitable variations that can occur in natural scenes such as partial occlusions, illumination variations, poses and appearance changes.

To cope with the problems mentioned above the tracker needs to be adaptive. Collins and Liu [1] first proposed a method to adaptively select color features that best discriminate the object from the current background. Lim et al. [2] used incremental subspace learning for tracker updating and Avidan [3] used an adaptive ensemble of classifiers. Grabner[4,5] have designed an on-line boosting classifier that selects features to discriminate the object from the background. The on-line boosting algorithm demonstrates excellent real-time performance on the tracking task but faces drifting as the key problem. Thinking error accumulation is the major reason of drifting, Semi-supervised [6] and re-detect method [7] has been proposed to cope with drifting problem by combine the tracking method with a detector.

* Corresponding author.

In this paper, we focus on the drifting problem in online boosting tracking, show three shortcomings of original online boosting tracking algorithm which may cause the drifting problem and seek a solution by using fragment based methods.

2 On-line Boosting Tracking with Fragment

The goal of On-line boosting tracking is to combine N weighted α_i selector $\mathbf{h}_i(\mathbf{x})$ into an adaptive strong classifier $\mathbf{H}(\mathbf{x})$, which can be refined with each new sample, to find the object location[4,5].

From the experiment and analysis, we find out that there are three shortcomings in the original online boosting tracking.(1)The voting weight of the first few selectors are too big so only few of the weak classifiers can affect the final strong classifier. If there are mistakes in the first few selectors, drift will occur. (2)The importance weight λ , which is used to compute error rate, then is used to compute the voting weight, is passed globally. But in fact, the variation only occur in local area of the target and only affect the neighborhood, not the whole area of the target. (3)The tracking results are not stable because the weak classifiers are selected randomly from the global feature pool. The algorithm may work very well this time but drift seriously next time.

In [5], the weak classifier with the lowest estimated error is selected; the corresponding voting weight and the importance weight are updated and passed to the next selector. The above steps are processed in global object area and the correct or wrong result of previous selector will affect the voting weight of next selectors. But we know that the result of a selector should only affect its neighborhood, not the global area, because usually only some parts of the object change significantly. So the computation of voting weight should be modified to a more reasonable way.

We combine the online boosting tracking with fragments. In each fragment we build a selector, importance weight are only passed in fragment, not in global area. The final strong classifier is computed by linear combination of all the selected weak classifiers in all fragments.

Each patch in the object area has one selector, and the final strong classifier is linearly combined by the selected weak classifiers with corresponding voting weights. Each selector has M weak classifiers (features) and selects the best one (with the lowest estimated error).Selected weak classifier can represent the state of its patch. Once one new frame come, the lowest estimated error e_n , corresponding voting weight α_n , and the passing importance weight λ_n are all changed correspondingly (Note: we pass the λ_n in patch n , not in the global area, so we have λ_n ,not only one λ). Different from method in [5], we use number of correctly classified classifiers to compute the passing importance weight λ_n .

$$e_n = \arg \min_m (e_{n,m}), \quad e_{n,m} = \frac{\lambda_{n,m}^w}{\lambda_{n,m}^c + \lambda_{n,m}^w} \tag{1}$$

$$\alpha_n = \frac{1}{2} \cdot \ln \left(\frac{1 - e_n}{e_n} \right) \quad (2) \quad \lambda_n = \frac{\text{CorrectNum}_n}{M} \cdot \alpha_n \tag{3}$$

We show the shots of tracking results of our Tracker, Frag Tracker, and OAB Tracker in Fig.1.

3 Conclusion and Future Work

A novel approach to overcome the drifting problem in online boosting tracking was presented. Our tracker combines on-line boosting and object fragment to adapt the tracker to local variance. We pass the importance weight through local patches, not globally as in OABTracker, and get better result. But when there are dramatically rotations and illumination changes, our method should be improved to work better.



Fig. 1. Screen shots of tracking results

Acknowledgments. This research has been supported by funding from National Natural Science Foundation of China (61202168, 61201234), and Key project in Science and technology pillar program of Tianjin(10ZCKFGX00400).

References

- [1] Collins, R., Liu, Y., Leordeanu, M.: Online selection of discriminative tracking features. IEEE Trans. PAMI 27(10), 1631–1643 (2005)
- [2] Lim, J., Ross, D., Lin, R., Yang, M.: Incremental learning for visual tracking. In: Saul, L.K., Weiss, Y., Bottou, L. (eds.) NIPS, vol. (17), pp. 793–800. MIT Press (2005)
- [3] Avidan, S.: Ensemble tracking. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 494–501 (2005)

- [4] Grabner, H., Bischof, H.: On-line boosting and vision. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 260–267 (2006)
- [5] Grabner, H., Grabner, M., Bischof, H.: Real-time tracking via on-line boosting. In: Proceedings of BMVC, vol. 1, pp. 47–56 (2006)
- [6] Grabner, H., Leistner, C., Bischof, H.: Semi-supervised On-Line Boosting for Robust Tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 234–247. Springer, Heidelberg (2008)
- [7] Kalal, Z., Matas, J., Mikolajczyk, K.: P-n learning: Bootstrapping binary classifiers by structural constraints. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 49–56 (2010)
- [8] Adam, A., Rivlin, E., Shimshoni, I.: Robust fragments-based tracking using the integral histogram. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 798–805 (2006)

Nonrigid Object Modelling and Visualization for Hepatic Surgery Planning in e-Health

Suhuai Luo¹ and Jiaming Li²

¹ The University of Newcastle, Australia

Suhuai.luo@newcastle.edu.au

² The CSIRO ICT Centre, Australia

Jiaming.li@csiro.au

Abstract. This paper introduces an automatic approach of nonrigid object modelling and visualization for hepatic surgery planning, in particular, for live donor liver transplantation and accurate liver resection for cancer in e-health application. The proposed approach can build a system that supports radiologists in data preparation and gives surgeons precise information for making optimal decisions. It provides 3D representation of liver parenchyma and vasculature, and 3D simulation of patient specific data. The system is realized in four major stages, including registration of multimodal images; segmentation of liver parenchyma; extraction of liver vessels; and modelling and visualization of liver parenchyma and vessels. The approach is unique in that it integrates advanced techniques such as machine learning algorithm with a knowledge base of the organ. The details of these stages are described along with experimental results and discussions of the advantages of the approach over other approaches.

Keywords: visualization, modeling, segmentation, machine learning, surgery planning.

1 Introduction

Computer aided surgical planning (CASP) uses computer technology to conduct pre-surgery planning and in-surgery guiding and intervention. It has been widely used in many fields such as neurosurgery, orthopaedic surgery, hepatic surgery, etc. [1]. A CASP system is typically composed of four components including image registration, object segmentation, object modeling, and simulation. The focus of this research is to build a system of CASP for liver where live donor liver transplantation and accurate liver resection for cancer will be performed.

There exist a few computer aided surgical planning systems for liver [2-3]. These systems can present the structures of various liver vessels, generate resection proposal, offer 3D visualization, provide surgical simulation with cutting. However, among these systems, the most serious problems are the accurate segmentation of liver from its surrounding organs and extraction of liver vasculature in CT images.

This paper introduces an automatic approach of nonrigid object modelling and visualization for hepatic surgery planning, in particular, for live donor liver

transplantation and accurate liver resection for cancer in e-health application. In section 2, we describe the system. Finally in section 3, we conclude by summarising the approach and pointing out possible future pursuits in the area.

2 The Approach

The proposed nonrigid object modelling and visualization for hepatic surgery planning is realized in four major stages (see Fig. 1), including (1) registration of multimodal images, (2) segmentation of liver parenchyma, (3) extraction of liver vessels, and (4) modelling and 3D visualization of liver parenchyma and vessels.

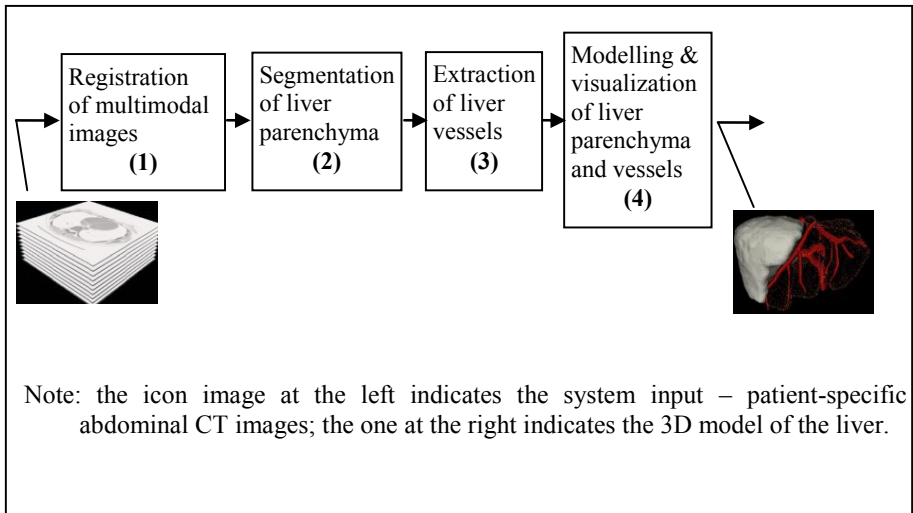


Fig. 1. Block diagram of the liver modeling and visualization for hepatic surgery planning system

(1) Registration of multimodal images

Image registration is the process of transforming different sets of images into one coordinate system so that these images can be compared or integrated as needed. In liver transplantation and resection, aligning all the CT image series to a global best fit, i.e., registration of multimodal images, is needed. Our registration algorithm performs optimal selections of the important components of registration, including transform, metric, mapper, optimizer, etc.

(2) Segmentation of liver parenchyma

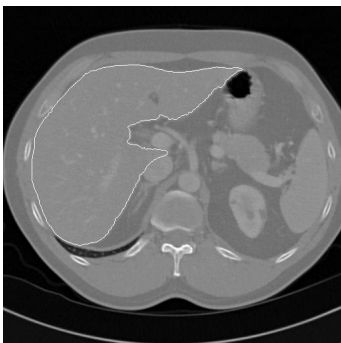
An advanced liver parenchyma segmentation algorithm is developed. It integrates texture analysis and machine learning with deformable surface model. The algorithm starts with wavelet-based [4] texture analysis on abdominal CT images to extract pixel level features. Then support vector machines (SVMs) [5] are used to classify

each pixel into either liver or non-liver. Finally an advanced 3D dynamic gradient vector flow (GVF) snake is developed to delineate the liver.

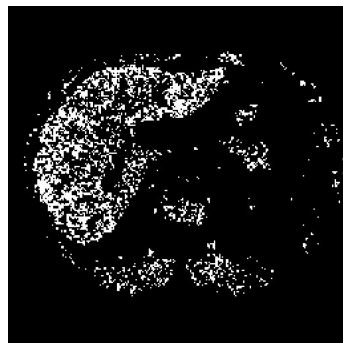
Fig. 2 presents an example of the output of SVM classifier. Left, is an original slice of abdominal CT image, where the liver is at the top-left corner, indicated with the white curve. Right, shows the output of SVM. From the figure, it can be seen that the SVM can classify most of the liver pixels correctly as a cluster of pixels.

(3) Extraction of liver vessels

A knowledge-based vasculature segmentation algorithm is developed to extract liver vessels. It segments liver vasculature in two major steps. Step one is mainly vasculature candidate calculation using local intensity distribution, where the knowledge of image properties is used to derive possible vascular pixels or voxels. Step two is a knowledge-based region growing process that is guided by the anatomy of the vasculature. This involves the selection of parameters for region growing including starting seeds, size of neighborhood, and resultant topology.



an original slice of abdominal CT image



corresponding output of SVM

Fig. 2. An example of the output of SVM classifier

(4) Modelling and visualization of liver parenchyma and vessels

After the liver parenchyma and vessels are well segmented, 3D modelling and visualization of liver is performed to provide surgeons with useful information for understanding complex liver anatomy. It is done in three major steps: forming an RGBA volume from the data, reconstruction of a continuous function from this discrete data set, and projecting it onto the 2D viewing plane (the output image) from the desired point of view.

3 Conclusions

This paper presents a system of computer aided surgical planning for liver where live donor liver transplantation and accurate liver resection for cancer is performed. The approach is unique in that it integrates advanced techniques such as machine learning algorithm with a knowledge base of the organ. Outputs of some stages, such as the

liver parenchyma segmentation, have demonstrated that the system can deliver an automatic solution of nonrigid object modelling and visualization for hepatic surgery planning. Possible further work is to provide facilities and tools for exploring data, specifying surgical path, checking donor eligibility, assessing organ condition, predicting graft volume, and monitoring donor and recipient outcome.

References

1. Peters, T., Cleary, K. (eds.): *Image-Guided Interventions: Technology and Applications*. Springer (2008)
2. Radtke, A., et al.: Computer-assisted surgery planning for complex liver resections: when is it helpful? A single-center experience over an 8-year period. *Ann. Surg.* 252(5), 876–883 (2010)
3. Selver, M.A., et al.: Patient oriented and robust automatic liver segmentation for pre-evaluation of liver transplantation. *Computers in Biology & Medicine* 38(7), 765–784 (2008)
4. Materka, A., Strzelecki, M.: *Texture analysis methods - a review*. Technical Report, Technical University of Lodz, Institute of Electronics (1998)
5. Cristianini, N., Shawer-Tatlor, J.: *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press (2005) ISBN 0521780195

Encoder/Decoder for Privacy Protection Video with Privacy Region Detection and Scrambling

Feng Dai, Dongming Zhang, and Jintao Li

Advanced Computing Research Laboratory, Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
{fdai, dmzhang, jtli}@ict.ac.cn

Abstract. Privacy region scrambling is an effective method to protect privacy information in videos. In this paper, we present an encoder/decoder system for privacy protection video. On the encoder side, the privacy region in video is automatically extracted and scrambled while encoding. On the decoder side, users can exactly restore the original video with a legitimate key otherwise only non-privacy part can be decoded correctly but the privacy regions are encrypted.

Keywords: private protection, video scrambling, privacy region detection.

1 Introduction

With the rapid development of information technology and people's widespread concern about public safety, video surveillance systems have penetrated into all aspects of our lives. However, incessant monitoring makes people begin to pay more attention to personal privacy. Privacy region scrambling is one of the major technologies for private protection video [1]. Generally, the scrambling process is driven by a key. Anyone without the key can only see non-privacy region with the privacy region scrambled. When necessary, descrambler can exactly restore the original video with a legitimate key.

Encoder of privacy protection video mainly involves three parts: privacy region extraction, privacy region scrambling and video coding. Face or motion detection methods are commonly used to extract private regions. And the detected region is scrambling either before video coding or during video coding.

2 Encoder for Privacy Protection Video

Fig.1 illustrates the framework of the proposed Encoder of privacy protection Video. There are mainly three parts of the system. First, by utilizing data generated during video encoding, the privacy region is extracted. Then the detected regions are directly protected by quantized coefficients scrambling. To prevent drift error, a coding restricted scheme is employed.

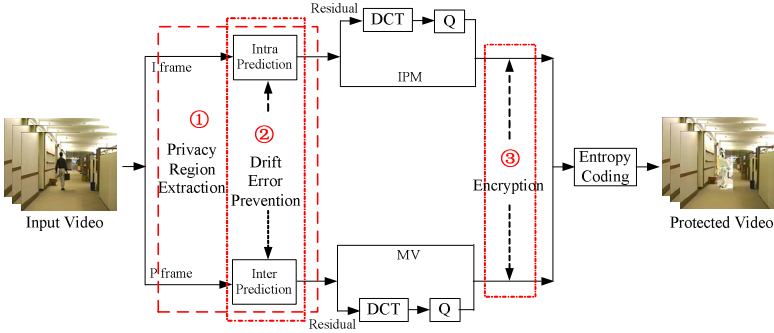


Fig. 1. Encoder for Privacy Protection Video

2.1 Motion Detection with Encoding Information

Motion is important information in video, which is critical for privacy protection. So moving objects are detected and protected in our system. In privacy protection video, video encoding is very time consuming [2] and all process must be finished in real time with additional motion detection. So fast motion detection must be adopted. Different from pixel domain detection and compressed domain detection, we proposed motion detection with encoding information. This method can extract moving objects directly during video encoding, so the computation time is reduced, which is attractive for real-time applications.

2.2 Transform Domain Scrambling

After obtaining the privacy regions, transform domain scrambling is applied to data related to these regions. The random sign inversion method is used in our system. In encoder, a pseudorandom number generator (PRNG) initialized by a key is used to produce random number sequences. Then the sign of quantized coefficients (defined as $qC[i]$, $i=0\dots15$) of each 4×4 block in privacy region is pseudo-randomly flipped for each i as follows:

$$qC[i] = \begin{cases} -qC[i] & \text{random_bits} = 1 \\ +qC[i] & \text{otherwise} \end{cases} \quad (1)$$

2.3 Drift Error Prevention

To improve coding efficiency while preventing drift error, mode restricted intra prediction (MRIP) and search window restricted motion estimation (SWRME) are proposed in our early work [3].

2.4 Decoder for Privacy Protection Video

Because the video bitstream is compliant with video coding standard, it can be displayed by a standard decoder but the privacy region scrambled. If we want to

decode the whole original image, the encoded video must be decoded by the specific video decoder. Anyone without the key can only see non-privacy data with the privacy region scrambled. Authorized users can exactly restore the original video with a legitimate key.

3 Demonstration

Fig.2 illustrates the encoder/decoder of private protection video user interface. Firstly, we choose a video to encrypt and input a key, then click the “encrypt” button, and the private region will be scrambled while the video encoding. Fig.2(a) shows the decoded image with the right key and Fig.2(b) shows the scrambled image with a wrong key.

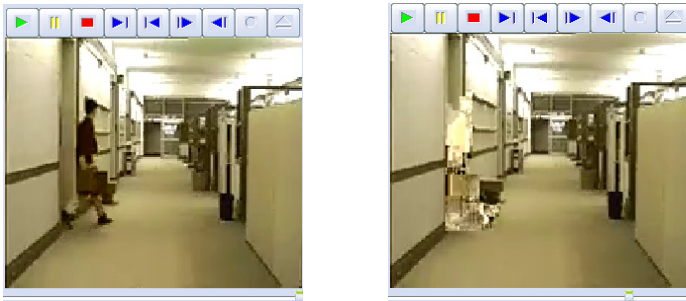


Fig. 2. (a) Decoded image with the right key (b) Scrambled image with a wrong key

Acknowledgments. This work is supported by National Nature Science Foundation of China (61102101, 61272323), National Key Technology Research and Development Program of China (2012BAH06B01).

References

1. Dufaux, F., Ebrahimi, T.: Scrambling for Privacy Protection in Video Surveillance Systems. *IEEE Trans. Circuits and Systems for Video Technology* 18, 1168–1174 (2008)
2. Zhang, Y., Yan, C., Dai, F., Ma, Y.: Efficient Parallel Framework for H.264/AVC Deblocking Filter on Many-core Platform. *IEEE Trans. on Multimedia* 14, 510–524 (2012)
3. Tong, L., Dai, F., Zhang, Y., Li, J.: Prediction restricted H.264/AVC video scrambling for privacy protection. *Institution of Engineering and Technology Electronics Letters* 46, 47–49 (2010)

TVEar: A TV-tagging System Based on Audio Fingerprint

Tao Jiang¹, Jiahong Li^{1,2}, Rihui Wu^{1,2}, and Kang Xiang^{1,2}

¹ Advanced Computing Research Laboratory, Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

² University of Chinese Academy of Sciences, Beijing, 100049, China
{jiangao01, lijiahong, wurihui, xiangkang}@ict.ac.cn

Abstract. This demo presents a TV-tagging system named TVEar based on audio fingerprint. It is a content-based audio information retrieval system, and has the ability to listen to a couple seconds of a TV show and determine what show is being watched. TVEar is robust to noisy environments, such as office/street/car environments. This system is designed to make a remarkable entry into social media.

Keywords: Audio fingerprint, TV-tagging, audio information retrieval.

1 Introduction

With the increase of smartphone's and table PC's popularity, more and more users keep their device in hand while watching TV. These device owners have the engage with content related to the TV, either by looking up information related to the show or looking for deals and general information on products advertised on TV. TV-tagging is new version of social media. People have a desire to share their idea about the TV show with their friends. They also have curiosity about what their friends watching now. Both the desire and the curiosity have a time limit, and would be lost soon. It is not a smart way to join a talk group corresponding to the TV show by a TV guide. There are so many TV channels that it may take several seconds to find the correct one from the directory. TV-tagging is also a new advertising mode and sales mode. The customers could get a promotion for tagging a selected TV advertisement. IntoNow[1] from Yahoo is a popular TV-tagging application, and it successfully connects the TV screen with second screen. TVEar provides the friendly user interface which is similar to the one of IntoNow.

The kernel of our product is a content-based audio information retrieval (AIR) [2, 3] system based on audio fingerprint. With this AIR system, TVEar can identify which TV channel is being watched. It also can recognize the music and movie playing during TV shows, down to the individual episode, within second. TVEar can recognize a show even if it's airing live for the first time.

TVEar is robust to noisy environments, such as office/street/car environments. The fingerprints of the audio signal would be masked or distorted in these noisy

environments. With a longer record to take more fingerprints, TVEar could also give the correct results.

2 System Overview

TVEar can identify both live TV channel and historical audio data which come from music, movie, episode or any other audio clips. Live audio data and historical one cannot be recognized in the same way. The number of TV channel is limit, and each of them only requires several minutes audio record. The search space for live audio data contains only several thousand seconds audio record which should keep fresh and be updated all the time, while that for historical audio data contains million minutes audio record which do not need a frequent update. For the enormous difference in search space and update frequency, the audio information retrieval of TVEar is divided into two retrieval platforms, as is shown in Figure 1. One of them is for the live TV channel recognition, and another one is for historical audio data. These two parts do not share the same audio feature or retrieval algorithm. The recognition procedures of these two parts are parallel. There is no prior knowledge of that the query audio clip comes from live TV channel or historical audio data.

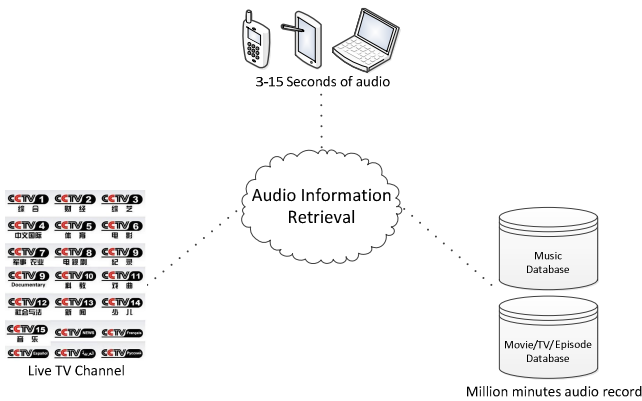


Fig. 1. Overview of TVEar

Due to limit search space and real-time update, the live TV channel recognition does not need an index. There is also not enough time to build it. In this recognition process, each channel keeps several minutes audio record as its template. The record is transformed to a time series feature vector, named channel-vector. The query audio clip is several seconds and also transformed to a vector, named query-vector. The query-vector slides on channel-vector, and the similarity between them is calculated frame by frame. The query clip has an offset with the channel record, which is shorter than half a frame. Two or more frames shift would lead to a low recall rate.

Due to enormous search space of historical audio data, an index is necessary to speed up its search. The index building needs the spare representation of audio

fingerprint. Wang A.L. provides a proper one[4] which is used in the popular music recognition application--shazam[5]. An inverted file implemented as a sorted array structure stores the list of audio fingerprint in a sorted array, including the ID of audio clips associated with each audio fingerprint and a link to the audio clip containing that audio fingerprint. The recognition process is similar to a search engine with a temporal rule. As is shown in Figure 1, music data and speech data, such as movie/TV show/episode, have their respective database. Music data carries richer information than speech data, and contains more audio fingerprints than it.

3 Demonstration

Figure 2 shows two examples of TVEar, which includes 20 live TV channels and 3 million music songs. The smart phone records an audio clip coming from TV. Then the clip is coded at a low rate and sent to TVEar. Two retrieval platforms recognize this clip independently, and both of them can return their own recognition result. (A) is a live TV channel recognition. (B) is a music recognition. The paly time to the music is also determined.

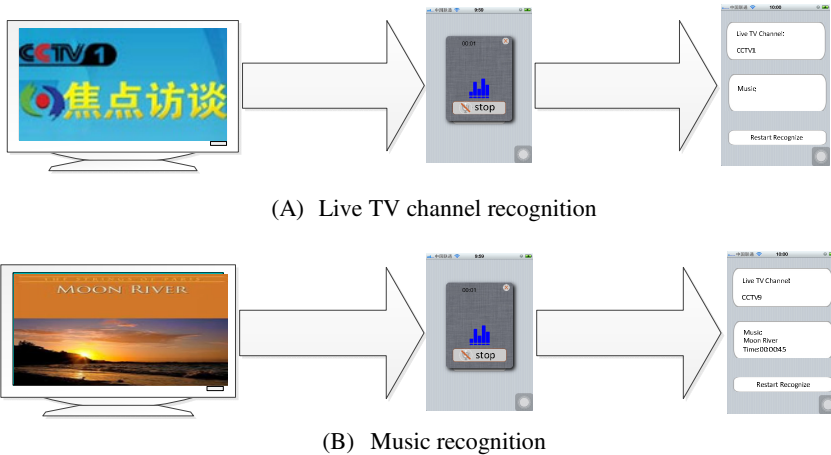


Fig. 2. Two examples of TVEar recognition

Acknowledgments. This work is supported by Co-building Program of Beijing Municipal Education Commission.

References

1. <http://www.intonow.com>
2. Wang, M., Ni, B., Hua, X.S., Chua, T.S.: Assistive Tagging: A Survey of Multimedia Tagging with Human-Computer Joint Exploration. *ACM Computing Surveys* 44, 25 (2012)

3. Casey, M.A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., Slaney, M.: Content-Based Music Information Retrieval: Current Directions and Future Challenges. *Proceedings of the IEEE Journal* 96, 668–696 (2008)
4. Wang, A.L.: An Industrial-Strength Audio Search Algorithm. In: 4th Symposium Conference on Music Information Retrieval, ISMIR 2003, pp. 7–13 (2003)
5. <http://www.shazam.com>

VTrans: A Distributed Video Transcoding Platform

Zhe Ouyang^{1,2}, Feng Dai¹, Junbo Guo¹, and Yongdong Zhang¹

¹ Advanced Computing Research Laboratory, Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China

² University of Chinese Academy of Sciences, Beijing, 100190, China
{ouyangzhe, fdai, guojunbo, zhyd}@ict.ac.cn

Abstract. This demo presents a distributed video transcoding platform named VTrans, which utilizes the technology of distributed video transcoding. It can realize the fast transcoding of videos. The fast video transcoding method used in this platform is a video GOP-level and slice-level combined parallel mode, which can accelerate the process of video transcoding in time and space respectively. By using the system, users' waiting time of transcoding a video is reduced, and the use ratio of system resource is enhanced.

1 Introduction

Along with the development of Internet technology, video has become an indispensable part of people's daily life, as an example, there are roughly 24 hours of new videos uploaded to YouTube every minute, and YouTube hits over a billion daily video views [1]. At the mean time, users have different requirements for video qualities, codecs and formats, etc. Especially, most Internet videos are PC oriented [2], not fit for mobile devices. Since a mass of demands of video transcoding exist, a kind of high performance video transcoding system is desired.

In this presentation, the VTrans transcoding platform can exactly meet the tremendous transcoding demands, it adopts distributed video transcoding technology in the background of the transcoding system, supporting various common video formats, which can realize rapid video transcoding and reduce the user's transcoding waiting time.

2 System Overview

2.1 System Architecture

Fig.1 shows the architecture of the VTrans transcoding platform, the system mainly contains four parts, namely video download module, video transcoding module, video management module and message notification module. The four modules are all data-driven, which communicate with each other through the message queue in the database, as a result, the four modules are fairly loose coupled. The process of that an

Internet video enters the system is as follows, video download → video transcoding → video information extraction → message notification, which is a 4 parts pipeline.

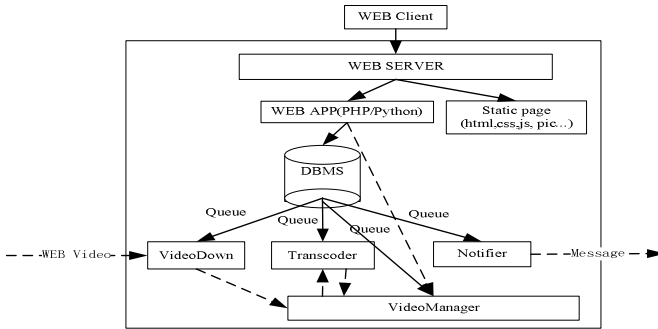


Fig. 1. The architecture of VTrans

2.2 Distributed Transcoding

The architecture of the video transcoding module is distributed. There is a controller and many transcoders, the controller fetches video transcoding tasks from DBMS, then analyses the video and divides the source task into some sub-tasks, after that, the transcoders get the sub-tasks from controller and do the transcoding work, after all the sub-tasks have been completed, a merge task must be done to combine the multiple video clips into a integrated video. The task splitting procedure is not trivial, many factors must be taken into account, such as the video duration, the codecs and resolutions of source videos and target videos, etc. For example, the hierarchical structure of H264 has six levels: sequence, groups of pictures (GOPs), pictures, slices, macroblocks (MBs) and blocks [3][4]. Theoretically, the parallel transcoding can be realized on the above six levels, but not all the codecs supporting the parallel transcoding on the all the six levels and for the sake of simplicity of realization, we adopt only the GOP level parallel transcoding, while, if the target codec is H264, just as Fig.2 shows, the GOP level and slice level parallel transcoding will be used together, and we will have a significant gain in transcoding speed.

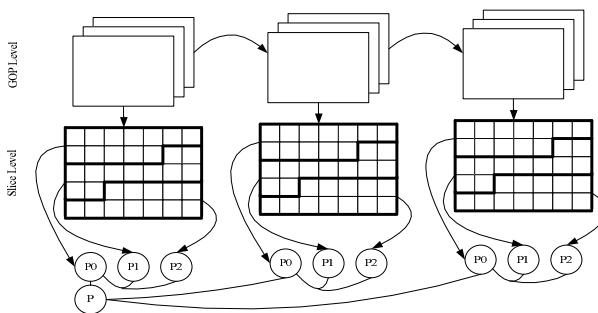


Fig. 2. GOP level and slice level combined parallel mode

3 Demonstration

Fig.3 shows the task management page of VTrans, we can create multiple video transcode tasks one time and have a centralized control of them. For example, we can cancel a running task or restart a stopped task. The statuses of tasks are clearly displayed, it's convenient for us to monitor the transcoding process and estimate the time remained.

Task ID	URL	Input URL	Input length	Output Num	Submitted Time	Status	Progress Bar	Operating
144	http://v.ifeng.com/ml/mainland/201209/2db696fc-12c4-4971-b58d-3406751f614e.shtml		--	1	2012-09-28 15:07:51	Downloading	69	Cancel Restart
143	http://v.ifeng.com/v/abjshl/index.shtml#f8b5f13b-59ce-4041-bebb-4aa8b7677a4a		--	1	2012-09-28 15:07:38	Download fail	0%	Restart
142	http://tv.sohu.com/20120926/n354008974.shtml#297		--	1	2012-09-28 15:07:28	Downloading	12	Cancel Restart
141	http://tv.sohu.com/20120920/n353592970.shtml		--	1	2012-09-28 15:07:18	Downloading	30	Cancel Restart
140	http://www.tudou.com/programs/view/X5Y22zIQSI0/?fr=rec2		00:02:05	1	2012-09-28 14:56:52	Transcoding Success	100	
139	http://www.tudou.com/programs/view/X5Y22zIQSI0/?fr=rec2		00:02:05	1	2012-09-28 14:55:04	Transcoding Success	100	
138	http://www.tudou.com/programs/view/X5Y22zIQSI0/?fr=rec2		00:02:05	1	2012-09-28 14:54:09	Transcoding Success	100	
137	http://www.tudou.com/programs/view/X5Y22zIQSI0/?fr=rec2		00:02:05	1	2012-09-28 14:54:09	Transcoding Success	100	
136	http://www.tudou.com/programs/view/X5Y22zIQSI0/?fr=rec2		00:02:05	1	2012-09-28 14:54:06	Transcoding Success	100	
135	http://www.tudou.com/programs/view/X5Y22zIQSI0/?fr=rec2		00:02:05	1	2012-09-28 11:59:18	Transcoding Success	100	
134	http://www.tudou.com/programs/view/X5Y22zIQSI0/?fr=rec2		00:02:05	1	2012-09-27 11:08:48	Transcoding Success	100	
133	http://v.ifeng.com/ml/mainland/201209/2db696fc-12c4-4971-b58d-3406751f614e.shtml		00:00:44	1	2012-09-27 11:07:57	Transcoding Success	100	

Fig. 3. Task management page of VTrans

Acknowledgments. This work is supported by National Nature Science Foundation of China (61102101, 61272323), National Key Technology Research and Development Program of China (2012BAH06B01), Co-building Program of Beijing Municipal Education Commission.

References

- Davidson, J., Liebold, B., Liu, J., Nandy, P.: The YouTube Video Recommendation System. In: RecSys 2010, Proceedings of the Fourth ACM Conference on Recommender Systems, pp. 293–296 (2010)
- Li, Z., Huang, Y., Liu, G., Wang, F., Zhang, Z.L., Dai, Y.: Cloud Transcoder: Bridging the Format and Resolution Gap between Internet Videos and Mobile Devices. In: Network and Operating System Support for Digital Audio and Video, NOSSDAV (2012)
- Franché, J.-F., Coulombe, S.: A Multi-Frame and Multi-Slice H.264 Parallel Video Encoding Approach with Simultaneous Encoding of Prediction Frames. In: Consumer Electronics, Communications and Networks (CECNet), pp. 3034–3038 (2012)
- Zhang, Y., Yan, C., Dai, F., Ma, Y.: Efficient Parallel Framework for H.264/AVC Deblocking Filter on Many-core Platform. IEEE Trans. on Multimedia, 510–524 (2012)

Fast ASA Modeling and Texturing Using Subgraph Isomorphism Detection Algorithm of Relational Model

Feng Xue^{1,2}, Xiaotao Wang¹, Feng Liang¹, and Pingping Yang¹

¹ School of Computer Science and Information, Hefei University of Tech., Hefei, China

² State Key Lab of Virtual Reality Tech. and System, Beihang University, Beijing, China
{iamxuefeng, wangxiaotaolhc, liang.feng366, iamyangping}@163.com

Abstract. In this paper, a new method based on Subgraph Isomorphism Detection (SID) algorithm was proposed to automatically construct Anhui-Styled Architecture(ASA) models. Firstly, by analyses intrinsic features of ASA, we setup architecture module database. Then use SID algorithm to get a topology graph and traverse each node of the topology graph. Finally, render these graph nodes to get 3D model of ASA.

Keywords: Anhui-Styled Architecture(ASA), module construction, SID, Texture.

1 Introduction

In recent years, automatically and rapidly modeling of large scene becomes a hot topic in the fields of virtual reality research. To deal with this problem, researchers have proposed many modeling algorithms, which can be divided into two categories: rule based modeling method[1,2] and parametric and modular based modeling method[3-5] which have made a lot of contributions in rapid 3D modeling. However, ASA models often have its highlighted architectural styles, and the constraints between different function modules are much stricter than those box-like city buildings, and most of previous methods are not applicant for ASA. To construct ASA building models quickly and automatically, we present a fast modeling method based on the SID algorithm.

2 Outline of Proposed Algorithm

We proposed an automatic modeling method which base on modules decomposition and subgraph isomorphism detection algorithm, as shown in Fig. 1

3 Rapid Modeling Using SID Algorithm of Relational Model

The construction of topology database of ASA is the base of our SID algorithm. In the database we store information of a graph and its subgraph, including node number of

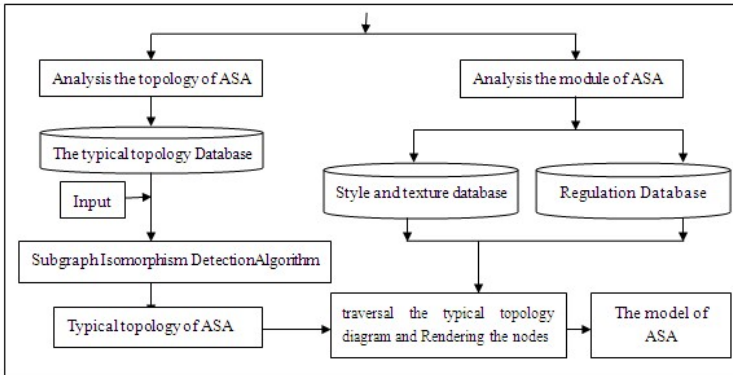


Fig. 1. Outline of the proposed algorithm

each subgraph, the degree number of each nodes, ids of each graph and subgraph etc. In order to make the algorithm more accurate, each subgraph we stored in database should have more than three nodes. This is because a subgraph that with little nodes (less than 3 nodes) may be matched by most of input graph.

There are many algorithms used for subgraph isomorphism detection, such as FG-Index, C-Tree and Graph Decomposition Index, but most of them are and time-consuming. To speed up graph search phase of modeling, we present a subgraph isomorphism algorithm based on Relational Graph Decomposition Index(RGDI).

In our SID algorithm, two graphs are supposed to be matched when: $DNa \geq DNi$ Where DNa is the degree number of a topology graph in database, and DNi represents the degree number of input topology graph.

4 Experimental Results

When 3D models generated by our SID algorithm are constructed, we use texture synthesis and texture mapping techniques to render some classic ASA style textures to the building model surfaces, as shown in Fig. 2, where individual building has different layout patterns, geometry shapes and texture styles from each other given different initializations and layout patterns. Like [6], we can make a video by linking each step of node module construction to show the dynamic inference process of SID.

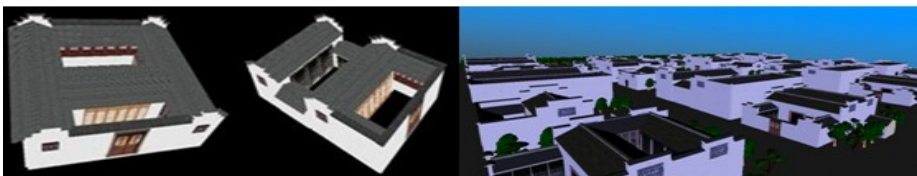


Fig. 2. Automatic modeling results of ASA with textures

5 Summary

In this paper, we analysis the structural feature and module decomposition of ASA, use SID algorithm to get a matched topology, traverse each node of the topology graph, construct the model of ASA automatically.

Acknowledgment. The work was supported by the Natural Science Foundation of China (No. 61202283), the Open Project of State Key Laboratory of Virtual Reality Technology and Systems of China (No. BUAA-VR-10KF-5).

References

1. Parish, Y.I.H., Müller, P.: Procedural modeling of cities. In: Proceedings of ACM SIGGRAPH Computer Graphics, Los Angeles, CA, USA. Annual Conference Series (2001)
2. Müeller, P., Zeng, G., Wonka, P., et al.: Image-based Procedural Modeling of Facades. *ACM Transactions on Graphics* 26(3), 85 (2007)
3. Müeller, P., Wonka, P., Haegler, D., et al.: Procedural modeling of buildings. *ACM Transactions on Graphics* 25(3), 614–623 (2006)
4. Ullrich, T., Settgast, V., Fellner, D.W.: Semantic fitting and reconstruction. *Journal on Computing and Cultural Heritage* 1(2), 560–574 (2008)
5. Liu, Y., Xu, C., Zhang, Q., et al.: Ontology based semantic modeling for Chinese ancient architectures. In: The Twenty-First National Conference on Artificial Intelligence (AAAI), pp. 16–20. AAAI, Boston (2006)
6. Wang, M., Hong, R., Yuan, X.-T., Yan, S., Chua, T.-S.: Movie2Comics: Towards a Lively Video Content Presentation. *IEEE Transactions on Multimedia* 14(3), 858–870 (2012)

An Approach for Browsing Video Collections in Media Production

Werner Bailer, Wolfgang Weiss, Christian Schober, and Georg Thallinger

JOANNEUM RESEARCH Forschungsgesellschaft mbH
DIGITAL – Institute for Information and Communication Technologies
Steyrergasse 17, 8010 Graz, Austria
`firstname.lastname@joanneum.at`

Abstract. This paper describes a video browsing tool for media (post-) production, enabling users to efficiently find relevant media items for redundant and sparsely annotated content collections. Users can iteratively cluster the content set by different features, and restrict the content set by selecting a subset of clusters. In addition, similarity search by different features is supported. Desktop and Web-based variants of the user interface, including temporal preview functionality, are available.

1 Introduction

The proposed video browsing tool attended the Video Browser Showdown (VBS) 2012 [3]. The tool has been designed for applications in (post-) production phase of movie and broadcast production. In this application scenario, users typically deal with large amounts of audiovisual material with a high degree of redundancy and need to select a small subset for use in a production. Newly shot material is typically sparsely annotated, thus the browsing tool has to rely on automatically extracted features. The content sets in this application are typically larger than those used in the Video Browser Showdown, reaching about 100 hours.

Automatic content analysis is performed during the ingest of content. Currently, camera motion estimation, visual activity estimation, extraction of global color features and estimation of object trajectories are performed. The extracted features are represented using the MPEG-7 Audiovisual Description Profile [1] and indexed in an SQLite database.

In order to select content, the user follows an iterative selection process, consisting of alternating steps of clustering and selecting subsets of the current data set. Users cluster content by one of the automatically extracted features and can then select relevant clusters to reduce the content set. Further clustering by the same or other features can then be applied to the reduced set. In addition, items similar to one of the items in the cluster can be retrieved. A more detailed description of the tool and the browsing process can be found in [2].

2 Browsing User Interface

The central component of the video browsing tool's user interface is a light table (cf. Figure 1). The light table shows the current content set and cluster structure

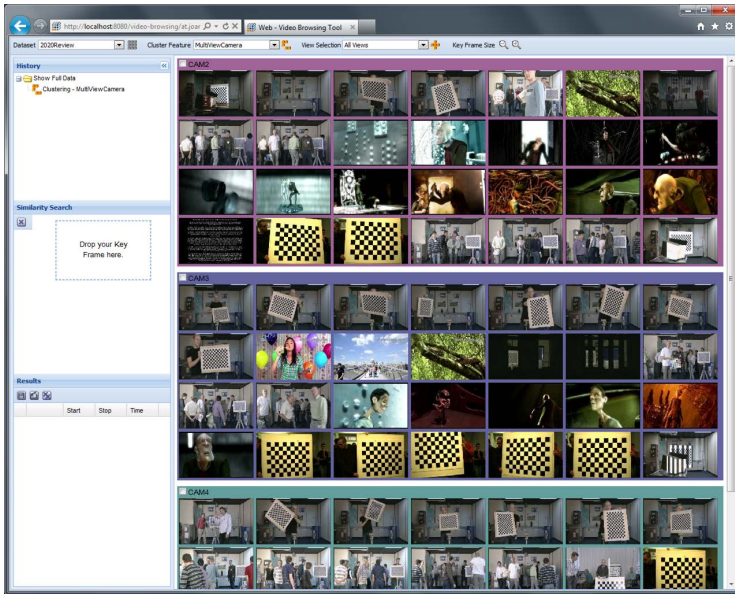


Fig. 1. Screenshot of the Web-based video browsing tool

using a number of representative key frames for each of the clusters. The clusters are visualized by colored areas around the images. The size of the images in the light table view can be changed dynamically so that the user can choose between the level of detail and the number of visible images without scrolling. By clicking on a key frame in the light table view, a video player is opened and plays the segment of the video that is represented by that image. The temporal context of a key frame is shown by a time line of temporally adjacent key frames that appears when the user moves the mouse over a frame. This time line shows one line of key frames which is limited by the width of the screen. If the user wants to get a broader range of temporally adjacent key frames, it is possible to zoom in. Therefore, a larger list of key frames is shown in the light table.

The tool provides support for performing similarity search based on following features: camera motion, motion activity, color layout, multi-view media item and multi-view camera. To execute a similarity search, the user drags a key frame into the similarity search area (see also Figure 2), selects a similarity search option and executes the search. In addition, segments from the temporal proximity of a result segment can be retrieved. The latter supports the user in cases where a presented item is topically similar to the wanted item and might thus be nearby in the programme, but is not similar in terms of visual features.

On the left side of the application window the history and the result list are displayed. The history window automatically records all clustering and selection actions done by the user. By clicking on one of the entries in the history, the user can go back to a previous point. Then users can choose to discard the subsequent steps and use other cluster/selection operations, or to branch

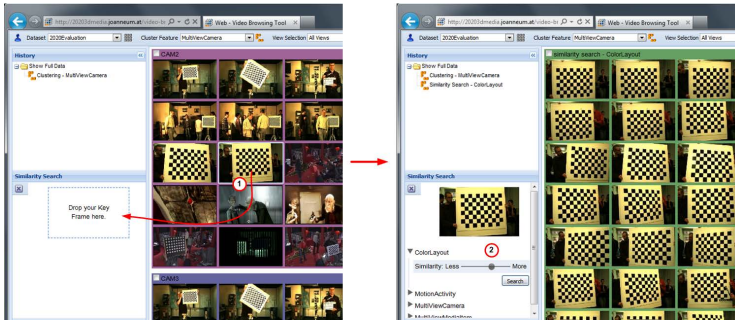


Fig. 2. Executing the similarity search

the browsing history and explore the content using alternative cluster features. The result list can be used to memorize video segments and to extract segments of videos for further video editing, e.g. as edit decision list (EDL). Users can drag relevant key frames into the result list at any time, thus adding the corresponding segment of the content to it.

The user interface is available both as desktop and as a Web-based version, offering the same functionality. Both use the same backend implementation, which is accessible as a SOAP Web service for the Web-based client.

Acknowledgements. The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreements n° 215475, “2020 3D Media – Spatial Sound and Vision” (<http://www.20203dmedia.eu/>) and n° 287532, “TOSCA-MP - Task-oriented search and content annotation for media production” (<http://www.tosca-mp.eu>).

References

1. Information technology - multimedia content description interface - part 9: Profiles and levels, amendment 1: Extensions to profiles and levels. ISO/IEC 15938-9:2005/Amd1:2012 (2012)
2. Bailer, W., Weiss, W., Kienast, G., Thallinger, G., Haas, W.: A video browsing tool for content management in post-production. *International Journal of Digital Multimedia Broadcasting* (March 2010)
3. Bailer, W., Weiss, W., Schober, C., Thallinger, G.: A Video Browsing Tool for Content Management in Media Post-Production. In: Schoeffmann, K., Merialdo, B., Hauptmann, A.G., Ngo, C.-W., Andreopoulos, Y., Breiteneder, C. (eds.) *MMM 2012*. LNCS, vol. 7131, pp. 658–659. Springer, Heidelberg (2012)

DCU at MMM 2013 Video Browser Showdown

David Scott, Jinlin Guo, Cathal Gurrin, Frank Hopfgartner,
Kevin McGuinness, Noel E. O'Connor, Alan F. Smeaton,
Yang Yang, and Zhenxing Zhang

Dublin City University
Glasnevin, Dublin 9, Ireland

{dscott, jguo, cgurrin, fhopfgartner, asmeaton, yyang, zzhang}@computing.dcu.ie,
{mcguinne, oconnorn}@eeng.dcu.ie

Abstract. This paper describes a handheld video browser that incorporates shot boundary detection, key frame extraction, semantic content analysis, key frame browsing, and similarity search.

Keywords: video browser showdown.

1 Introduction

In last year's MMM Video Browser Showdown we participated with a handheld based video browsing system that supported two methods of search: concept search and key frame similarity [1]. Building on the experience that we gained while participating in this event, we compete in the next showdown with a more advanced browsing system. This paper provides a short overview of the features and functionality of our new system.

2 Data Processing

2.1 Video Segmentation

To ease access to the video data, we segment the video into its shots following Pickering et al. [2]. To avoid including similar frames, we remove duplicates by comparing the global color layout of all key frames within each shot.

2.2 Visual Concept Classification

An important feature of our browsing system is the ability to filter search results based on the appearance of certain concepts such as *persons*, *vehicles*, *on screen text*, or *landscapes*. This filtering approach, also referred to as concept-based search, is an effective method to bridge the semantic gap between low-level features and high-level semantics. We trained a set of concept detectors using the popular Bag-of-Visual-Word (BoVW) feature representation and Support Vector Machine (SVM) classifiers.

2.3 Visual Similarity Search

Similarity search is implemented on the key frames by aggregating local descriptors to form a low dimensional global image descriptor. We use sparse SIFT [3] as the local descriptor and perform aggregation using the VLAD method proposed in [4]. This method is a simplification of the Fisher kernel representation [5], but is lower dimensional and less costly to compute. VLAD descriptors will be computed by first performing k-means clustering (with $k = 64$) on the entire set of SIFT descriptors extracted for each video, then aggregating the differences between each local descriptor in an image and its nearest neighbor, and finally concatenating these local differences to form a $d \times k$ descriptor, where $d = 128$ is the dimension of the local SIFT descriptors. PCA is then used to reduce the dimension of the VLAD descriptors to 128 dimensions. Since this representation allows us to represent all the key frames a video in less than 50MB of memory, we omit the product quantization step and perform search by exhaustively computing the distance between the query VLAD vector for the image and all other key frames in the dataset, rather than using approximate nearest neighbors [6].

2.4 Visual Similarity-Based Clustering

To reduce the cognitive load associated with scanning through a large number of heterogeneous key frames, we implemented functionality that allows the user to group together visually similar key frames and browse the resulting groups. Grouping is achieved by performing agglomerative clustering on the VLAD descriptors extracted for similarity search.

2.5 Face Browsing and Search

We anticipate that providing functionality to allow users to get a high-level overview of all the human faces appearing in a video will be useful for queries involving people. To this end, we provide a face view that shows all the faces found in the video, and allow users to quickly navigate to the locations in the video in which selected faces appear. We use the Viola-Jones face detector [7] to first locate faces in the videos, and then to cluster these faces by using agglomerative techniques. This clustering also allows face-based search to be easily implemented: when the user chooses to search for similar faces on a given key frame, all images associated with the clusters containing any faces that appear in the key frame are retrieved and displayed.

3 Graphical User Interface

Figure 1 depicts a screenshot of our graphical user interface. It is designed to be used on a tablet PC: either an iPad or any Android tablet. As can be seen in the screenshot, clustered results (shots) are represented by their key frame. On the top of the interface, users can apply different filters and enable similarity search

in the title bar. After tapping (or clicking) on one of the key frames in the result list, a content menu appears and the user has the choice to (a) start playing the video shot, (b) find similar shots in the whole video, or (c) find shots containing similar faces.

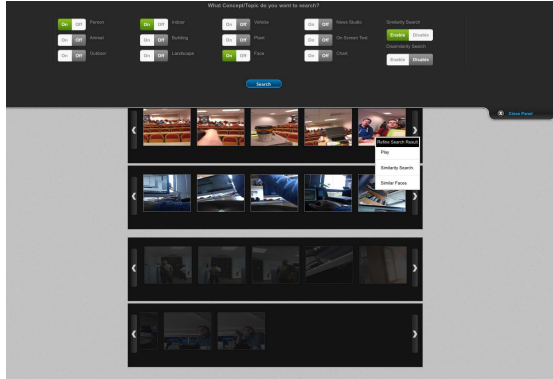


Fig. 1. Screenshot of the Video Browsing Interface

Acknowledgments. This research was supported by the Norwegian Research Council (CRI number: 174867), Science Foundation Ireland under Grant No.: 07/CE/I1147 and by the EU FP7 Project AXES ICT-269980.

References

1. Scott, D., Guo, J., Wang, H., Yang, Y., Hopfgartner, F., Gurrin, C.: Clipboard: A Visual Search and Browsing Engine for Tablet and PC. In: Schoeffmann, K., Merialdo, B., Hauptmann, A.G., Ngo, C.-W., Andreopoulos, Y., Breiteneder, C. (eds.) MMM 2012. LNCS, vol. 7131, pp. 646–648. Springer, Heidelberg (2012)
2. Pickering, M.J., R uger, S.M.: Evaluation of key frame-based retrieval techniques for video. *Computer Vision and Image Understanding* 92(2-3), 217–235 (2003)
3. Lowe, D.: Object recognition from local scale-invariant features. In: *IEEE International Conference on Computer Vision*, pp. 1150–1157 (1999)
4. J gou, H., Douze, M., Schmid, C., Perez, P.: Aggregating local descriptors into a compact image representation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3304–3311 (2010)
5. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1-8 (2007)
6. J gou, H., Douze, M., Schmid, C.: Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 117–128 (2011)
7. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001)*, vol. 1, pp. 511–518 (2001)

AAU Video Browser with Augmented Navigation Bars

Manfred Del Fabro, Bernd Münzer, and Laszlo Böszörményi

ITEC – Information Technology, Alpen-Adria-Universität Klagenfurt, Austria
{manfred,bernd,laszlo}@itec.aau.at

Abstract. We present an improved version of last year’s winner of the Video Browser Showdown. In a preprocessing step video segments are detected and clustered in several latent classes of similar content based on color and motion information. The navigation bars of our video browser are then augmented with different colors indicating where elements of the detected clusters are located. As humans are able to classify the content of clusters fast, they can benefit from this information when browsing through a video.

1 Introduction

The Video Browser Showdown (VBS) 2012 showed that certain Known-Item-Search (KIS) tasks can be performed effectively and efficiently with our AAU Video Browser[1]. It is solely based on intelligent interaction means and refrains from content analysis, but it takes use of the human abilities to recognize and classify items very fast. Therefore, scenes that are significantly different from the other scenes in a video or scenes that are expected at certain locations of a video (e.g. in most cases weather reports are at the end of news videos) can be found fast. On the other hand, specific tasks exist where it is difficult to find the searched scene with the same approach. In particular, in videos that show similar content from the beginning to the end (e.g. TV shows) or videos that consist of repeating similar situations (e.g. videos where anchorpersons or some sports acts are shown again and again) scenes are hard to find only by human observation.

This year we present an improved version of our video browsing tool, which combines automatic content analysis with human cognition to overcome the problems mentioned above. We present a video browsing tool that augments the navigation bars with additional information that indicates where certain content classes are located. The users still have to look for the searched items manually, but the search effort can be reduced by taking advantage of the additional information provided by the augmented navigation bars.

2 Augmented Navigation Bars

Augmented navigation bars are annotated with colored blocks, where each color represents a certain class of video content. An example is shown in Figure 1.

The idea is (1) to automatically detect repeating segments within videos, (2) to cluster them in groups of similar content and (3) to annotate the navigation bars with colors indicating the location of the clusters in the video. Each cluster is visualized by a different color. A large number of analysis methods can be used for step 1. For this approach we only rely on the color information and the motion information, which are extracted from the videos in order to cluster video segments into groups of similar color or motion. A latent indexing of the content is performed, thus no predefined classes are used.

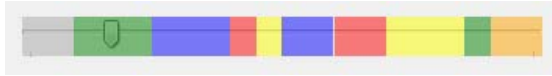


Fig. 1. Example of an augmented navigation bar

The classification of the emerging clusters is the part where the user comes into the loop. A preview panel is provided to the users, which shows the cluster centers of each cluster. Each preview is surrounded by a border colored with the same color that is used for marking all segments of the corresponding cluster on the navigation bar. The preview panel is shown next to the video windows in Figure 2. It helps the users to quickly make a basic discrimination of the content of a video. If the representative frames are not discriminative enough, users can load all segments that belong to a cluster in an own playlist by clicking on the cluster center in the preview panel. The elements of a playlist are ordered chronologically, thus scanning the items of a playlist from the beginning to the end is also an option to search for a certain video segment. An example of a playlist can be seen in the right part of the window in Figure 2.

The most important point regarding the presented video browsing application is that users are still interacting with videos and not only with static key-frames. Therefore, they experience videos in the same way as usual, but in addition the amount of time needed for certain KIS tasks can be reduced.

3 Conclusion

We improved our AAU Video Browser for the Video Browser Showdown 2013 with augmented navigation bars. The idea is to reduce the search space for users by coloring the navigation bars with different colors that indicate different clusters of video content. At the competition we want to investigate whether the problems that occurred in videos with recurring similar scenes can be overcome with this new approach.

The combination of well-known video interaction concepts with the results of content analysis has several advantages. The presented solution preprocesses the content and suggests classes of similar content to the users, thus they get a first

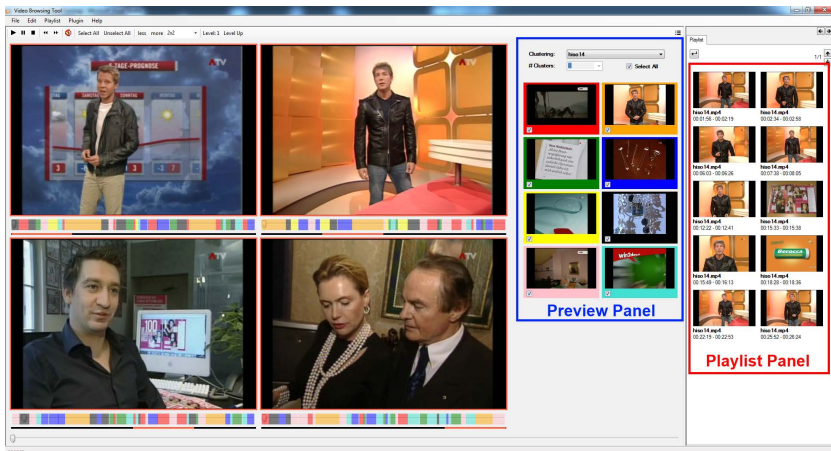


Fig. 2. The four windows at the left side can be used to browse four parts of a video in parallel. Next to them the preview panel is displayed. At the right side the playlist view is showing the segments of one cluster.

overview of the content of a video. Moreover, users can interact with videos as they were always used to. Compared to other key-frame-based tools they can still watch and interact with a video stream. The augmented information is only an additional help for them. Furthermore, the temporal order of the content gets preserved, thus users always have an overview of the temporal correlation of different segments.

We are going to investigate the integration of further analysis methods, such as local feature descriptors or face detection methods, and also the combination of different methods in future.

Acknowledgment. This work was supported by Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214 22573 33955.

Reference

1. Del Fabro, M., Böszörmenyi, L.: AAU Video Browser: Non-Sequential Hierarchical Video Browsing without Content Analysis. In: Schoeffmann, K., Merialdo, B., Hauptmann, A.G., Ngo, C.-W., Andreopoulos, Y., Breiteneder, C. (eds.) MMM 2012. LNCS, vol. 7131, pp. 639–641. Springer, Heidelberg (2012)

NII-UIT-VBS: A Video Browsing Tool for Known Item Search

Duy-Dinh Le¹, Vu Lam⁴, Thanh Duc Ngo², Vinh Quang Tran⁴,
Vu Hoang Nguyen³, Duc Anh Duong³, and Shin'ichi Satoh¹

¹ National Institute of Informatics,
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
{leddy, satoh}@nii.ac.jp

² The Graduate University for Advanced Studies
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan
ndthanh@nii.ac.jp

³ The University of Information Technology,
KM 20, Xa Lo Ha Noi, Thu Duc Dist, HCM City, Vietnam
{ducda, vnh}@uit.edu.vn

⁴ University of Science
227 Nguyen Van Cu, Dist. 5, HCM City, Vietnam
{lquv, tqvinh}@fit.hcmus.edu.vn

Abstract. This paper introduces a video browsing tool for the known item search task. The key idea is to reduce the number of segments to further investigate by several ways such as applying visual filters and skimming representative keyframes. The user interface is optimally designed so as to reduce unnecessary navigations. Furthermore, a coarse-to-fine based approach is employed to quickly find the target clip.

1 Introduction

The rapid explosion of video databases requires efficient tools for management and searching contents. Therefore, a tool for searching known items that a user has seen before has many potential applications. In the context of the video browser showdown competition, video browsing tools are evaluated through how well searchers interactively find a known video clip in one-hour video within a limited time. Building such a tool is challenging due to the following reasons:

- Time limit: The number of frames for one hour video is large, approx. 90,000 frames, while the allowed time is only 2 mins. Meanwhile, techniques for selecting candidates using visual information such as clustering and concept detectors are still not matured, and usually take time to operate, especially when the number of concepts and clusters are large.
- Space limit: The popular screen resolution is 1,024x768, while the minimum size of a thumbnail should be around 50x38 pixels to capture the content. Hence the maximum number of keyframes to be displayed is limited to approx. 300-400. This number is still much smaller than the total number of frames in one hour video.

We propose a system that helps to handle such problems. Specifically, the key ideas are:

- Coarse-to-fine based approach: At the coarsest level, one keyframe is used to represent content of each 2 minute-segment. By looking at these keyframes, some irrelevant segments can be eliminated quickly. At the finest level, 5 keyframes are used to represent content of each 10 second-segment. By looking at these keyframes, the segment that is relevant to the target clip can be identified and confirmed.
- Using filters to reduce the number of segments to be judged by users: They include a category-based filter that applies pre-trained classifiers to classify a keyframe into categories such as *Music*, *Entertainment*, *Indoor*, *Outdoor*, *Daytime*, and *Nighttime*; a layout based filter that select keyframes having similar layout with pre-defined layout patterns, and a color based filter that selects keyframes based on color distributions.

2 Framework Overview

2.1 Pre-processing Steps

The input one hour video is decomposed into short segments of 10 second length. The total number of such segments is 360. For each segment, 5 keyframes are selected to represent the visual content. These keyframes are grouped into clusters using a hierarchical clustering method.

For the coarsest level, the 360 segments are grouped into 30 super-segments. Each super-segment consists of 12 consecutive 10-second segments and is represented by one keyframe that is selected from the keyframes of the 10-second segments belonging to that super-segment.

2.2 Filters Using Visual Content

We use KAORI-SECODE [1] to build classifiers for annotation each keyframe with categories such as *Music*, *Entertainment*, *Indoor*, *Outdoor*, *Daytime*, and *Nighttime*. In addition to annotate keyframes with pre-define categories, we also annotate the keyframes with pre-defined color distributions, and layout patterns.

3 Video Browser Showdown Use-Case

The user interface of our browsing tool for the video browser showdown task is shown in Figure 1.

The top panel shows three filters including category-based filter, color-based filter, and layout-based filter. Using these filters, a number of relevant segments are selected to add to the list of candidates for further investigation. The right most position of the panel is a video player to play selected segments in fast forward mode (2x).

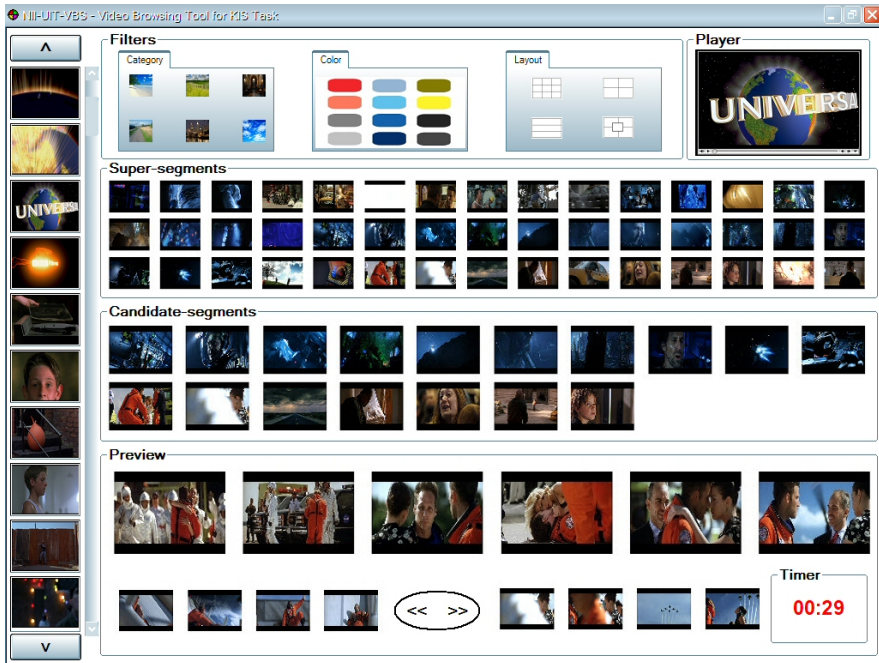


Fig. 1. The user interface of NII-UIT-VBS video browsing tool

The next panel (Super-segments panel) shows keyframes that represent for super-segments. Users will match these keyframes with the target clip content to select a subset of super-segments to add to the list of candidates.

The middle panel (Candidate-segments panel) shows segments that have been selected either manually by users or automatically by applying the filters.

The left panel shows segments that are randomly picked from the initial set of all segments. If one segment of interest is found, it can be added to the list of candidates.

By clicking on any segment in the list of candidates shown in the middle panel or the left panel, users will perform further investigation to check whether it matches the target clip.

The bottom panel (Preview panel) shows the finest level of the segment selected in the candidate list. The first row of the panel shows keyframes of the selected segments and the lower row shows keyframes of segments that are adjacent to the selected segment.

Reference

1. Le, D.-D., Satoh, S.: A Comprehensive Study of Feature Representations for Semantic Concept Detection. In: Proc. ICSC, pp. 235–238 (September 2011)

VideoCycle: User-Friendly Navigation by Similarity in Video Databases

Christian Frisson¹, Stéphane Dupont¹, Alexis Moinet¹,
Cécile Picard-Limpens¹, Thierry Ravet¹, Xavier Siebert², and Thierry Dutoit¹

¹ Université de Mons, TCTS lab, Boulevard Dolez 31, B-7000 Mons, Belgium

² Université de Mons, MathRO lab, Rue de Houdain 9, B-7000 Mons, Belgium
numediart Institute, Boulevard Dolez 31, B-7000 Mons, Belgium

Abstract. VideoCycle is a candidate application for this second Video Browser Showdown challenge. VideoCycle allows interactive intra-video and inter-shot navigation with dedicated gestural controllers. MediaCycle, the framework it is built upon, provides media organization by similarity, with a modular architecture enabling most of its workflow to be performed by plugins: feature extraction, clustering, segmentation, summarization, intra-media and inter-segment visualization. MediaCycle focuses on user experience with user interfaces that can be tailored to specific use cases.

1 The MediaCycle Framework

The MediaCycle framework has been developed at the numediart Institute¹ of the University of Mons since 2008. Its scope is to provide tools to create applications for the navigation by content-based similarity into multimedia databases, currently audio, image, video, text, sensor signals and files. It focuses “by design” on fostering new artistic practices (the numediart Institute aims at improving new media arts technologies) and stresses on providing a tailored user experience.

MediaCycle yields to modularity through a plugin architecture, from media types to all the steps of the workflow of organization by similarity: feature extraction, clustering, segmentation, summarization, visualization. Similarly to tools for computer-aided design, the generic graphical user interface of MediaCycle provides a main “canvas-like” space with browser and timeline views, and control panels associated to plugins. Some panels are generated automatically from plugin parameter serialization. We will describe in the following subsections all steps of the workflow of organization by similarity as offered by MediaCycle.

1.1 Audio and Video Feature Extraction

MediaCycle provides several algorithms for the extraction of low-level features. For audio signals, wrapper plugins around the YAAFE library and the Vamp

¹ MediaCycle:

<http://www.mediacycle.org> and numediart: <http://www.numediart.org>

audio analysis plugin system² allow access to verified and acknowledged algorithms from the music information retrieval and audio signal processing communities.

We initially developed features specific to dance videos [1], aside generic features such as: global motion orientation, optical flow, blob pixels speed, color moments; all based on OpenCV³ Time-dependent features are summarized by the calculation of their mean and standard deviation.

1.2 Clustering and Neighborhoods

MediaCycle proposes two modes of navigation: “clusters”, where the whole database is displayed on the screen, and “neighbors” where a user-defined small scale number of nearest neighbors are presented to the user. For clustering, we have been using K-Means. MediaCycle suggests neighbors using either an Euclidian distance or Pareto ranking, as explained in [1].

1.3 Segmentation

MediaCycle provides several segmentation methods, based respectively on:

1. the self-similarity matrix to compute a signal of novelty (see [2]);
2. the Bayesian Information Criterion (BIC) with two variants : browsing the data frame-by-frame or through a “divide-and-conquer” approach (see [2]);
3. a third-party library, Johan Mathé’s shotdetect⁴, segmenting by shots using consecutive frame pixel-by-pixel threshold-based color comparison.

Methods 1 and 2 are media-agnostic, 3 requires only video content. Once pre-computed, segmentation profiles associated to each method can be applied on the fly and affect the visualization content.

1.4 Visualization: Browser and Timeline

MediaCycle offers two views for navigation into media content: the “browser” positions media nodes in a 2D space for inter-media navigation, the “timeline” is dedicated to intra-media navigation. Several visualization techniques can be chosen for the browser: in clustering mode, the “propeller” tries to spread efficiently a user-definable number of clusters, a scatter plot (one user-definable feature dimension per x and y axes), a polar representation (one on radius, another on angle); in neighbors mode, radial and flat tree views show hierarchical nearest neighbors of given media elements. A fisheye distortion highlights segments hovered in any of both views with propagation to the other view.

² YAAFE:

<http://yaafe.sourceforge.net> and VAMP: <http://www.vamp-plugins.org>

³ OpenCV: <http://opencv.org>

⁴ shotdetect: <https://github.com/johmathe/Shotdetect>

1.5 Interaction and Devices

MediaCycle allows server-client application through the Transmission Control Protocol (TCP) [1] and external control through the OpenSoundControl (OSC) protocol [3] extended to TUIO, a protocol for Tangible User Interfaces. In short, MediaCycle applications can be displayed on touch-screen interfaces [1] and controlled via USB/HID devices such as jog wheels and multi-touch trackpads [3].

2 Our Approach to Known-Item Search with VideoCycle

We believe that our user interface VideoCycle built upon the MediaCycle framework can be used to browse single video files for known-item search as follows:

- Once the video file is imported (features are extracted and the video file is segmented), segments of the video are displayed in the browser view and organized by similarity, the user can choose the weights of the features to base the similarity on (here by checkboxes) and the positioning methods.
- The timeline view allows to navigate inside the imported video file, with a video player playing back the file, a “slider” view with an overlaid visualization technique that the user can choose among several (keyframes evenly-distributed in time, slit-scan, none), and a focus slider whose time interval is determined by the span of the aforementioned slider.
- A trackpad offers translation/rotation/zoom/reset gestures for navigation in the browser and skip/swipe gestures for the timeline. We pair it to a jog wheel, whose dial scrubs the timeline precisely and spring-loaded wheel adjusts the playback speed. Its keys are assigned to: previous/next segment, cue in/out and send the found “known item”, select features and visualizations.



Fig. 1. *The Remote Controller* (2003) by People Like Us alias Vicki Bennett in VideoCycle: zooming the inter-segment “browser” (up), scrubbing the intra-media “timeline” (down) with the jog shuttle wheel (down). Yellow dotted lines/arrows are annotations.

References

1. Tardieu, D., Siebert, X., Mazzarino, B., Chessini, R., Dubois, J., Dupont, S., Varni, G., Visentin, A.: Browsing a dance video collection: dance analysis and interface design. *Journal on Multimodal User Interfaces* 4(1), 37–46 (2010)
2. Dupont, S., Frisson, C., Urbain, J., Mahmoudi, S., Siebert, X.: Mediablender: Interactive multimedia segmentation. In: Dutoit, T. (ed.) QPSR of the Numediart Research Program, vol. 4, pp. 1–6 (2011), <http://www.numediart.org>
3. Frisson, C., Dupont, S., Siebert, X., Tardieu, D., Dutoit, T., Macq, B.: DeviceCycle: rapid and reusable prototyping of gestural interfaces, applied to audio browsing by similarity. In: *Proc. of the New Interfaces for Musical Expression, NIME* (2010)

Interactive Video Retrieval Using Combination of Semantic Index and Instance Search

Hongliang Bai¹, Lezi Wang², Yuan Dong¹, and Kun Tao¹

¹ France Telecom Research & Development - Beijing, 100190, P.R. China

² Beijing University of Posts and Telecommunications, 100876, P.R. China
{hongliang.bai,yuan.dong,kun.tao}@orange.com,
wanglezi@bupt.edu.cn

Abstract. We present our efficient implementation of interactive video search tool for Known Item Search(KIS) using the combination of Semantic Indexing(SIN) and Instance Search(INS). The interaction way allows users to index a video clip via their knowledge of visual content. Our system offers users a set of concepts and SIN module returns candidate keyframes based on users selection of concepts. Users choose keyframes which contains the interest items, and the INS module recommends frames with similar content to the target clip. Finally, the precise time stamps of the clip are given by the Temporal Refinement(TR).

1 Introduction

The task of KIS for Video Browser Showdown (VBS) is interpreted that users can interactively find the specific video clip in the database. This problem becomes a crucial issue due to exponential growth of video data.

VBS simulates a real-world KIS scenario. Users look for an item in the video collection that is known to be there, and the users have knowledge about the

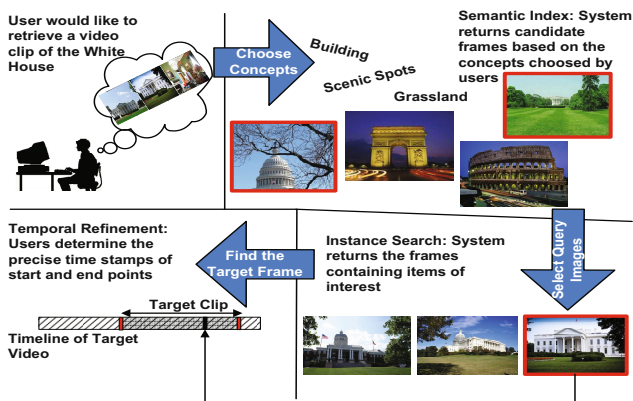


Fig. 1. Real scenario of indexing a video clip about "White House"

content. In the task, users are shown the video clip so that they have a good idea of what should be looked for. Moreover, there is no any metadata. Therefore automatic matching an not be performed. The aim of VBS is to evaluate interactive search systems.

2 System Framework

Our approach showcases that we solve the real-world KIS problem by combination of SIN and INS, as shown in Fig.1. Several concepts are pre-learned though visual information of training data. Users select corresponding concepts to retrieve a subset of candidates keyframes based on their knowledge about the target video clip. Then, they decide several images as the query of instance search. INS section contains a search loop so that users can refine the searching results. It can be neglected if users can detect the target keyframes in SIN results. TR is enabled once one target keyframe is retrieved, which decides the precise time stamps of the clip. The details of each module in our system are described in the next section

This section presents the details of our approach for real-world KIS problem. The system framework is illustrated in Fig.2. The keyframe extraction and analysis module is first described, and then we give details about SIN, INS and TR modules.

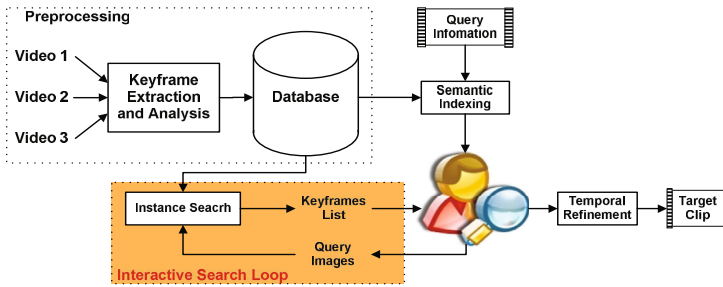


Fig. 2. Overview of our interactive video search system

2.1 Frame Extraction and Analysis

Processing all frames from the database video would be costly and inefficient. Thus, the frames are sub-sampled temporally. The clip of interest doesn't necessarily start and stop at shot boundaries. Therefore, in our system, we choose uniform sampling approach with 3 frames per second. Different visual features are used for frame analysis in SIN and INS module, as listed below:

SIN: Colors Coherence Vector (CCV), Local Binary Patterns (LBP), Histogram of Oriented Gradients (HOG), Dense Color SIFT, SIFT.

INS: Harlap and Maximally Stable Extremal Regions (MSER).

2.2 Semantic Index

Details of SIN algorithm can be referenced to our previous study of TRECVID Semantic Index task[1]. In the SIN module, 346-concept models are pre-learned through composite-kernel Support Vector Machine(SVM) and training data of TRECVID SIN 2011 and 2012. Users can chose several concepts based on their knowledge of the video content. And our system returns reference keyframes labeled with corresponding concepts. Nowadays, it still lacks of an effective algorithm to bridge the gap between low-level feature and high-level semantic. We observed that results of sematic labels indexing contain a lot of non-target keyframes, inconvenient for users to pick up the target frames. Thus, we decide to adopt the SIN module to get the initial results and introduce INS to optimize the searching. The INS module is described in the next section. In addition, if the target keyframes are detected in the SIN module, users can skip the INS module and enter TR.

2.3 Interactive Instance Search

The aim of INS module is detecting the frames which contains the items of interest. We adopt development data of TRECVID INS 2011 to build a Vocabulary Tree(VT) with 100 branches and 3 depth[2]. Then all reference keyframes are represented by projection of their features into the VT. Several images picked up by users in SIN module are regarded as the query of INS. The similarity between the tree projections of query images and reference frames are calculated. Fifty candidate reference keyframes with the highest similarity are shown to users after the spatial verification and query expansion. The INS section includes a random walk-based relevance feedback loop so that users can decide whether the every indexed frame is relevant or not[3].

2.4 Temporal Refinement

The temporal refinement is enabled once users find one target keyframe. The database video containing the target keyframe is shown to users. Based on the time stamps of target frame, users can check other nearby frames on the timeline. The target clip is retrieved when users determine the time stamps of start and end point.

References

- [1] Tao, K., et al.: The France Telecom Orange Labs (Beijing) Video Semantic Indexing Systems TRECVID 2011 (2011), <http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html>
- [2] Nistér, D., Stewénius, H.: Scalable Recognition with a Vocabulary Tree. In: CVPR (2006)
- [3] Rota Bulò, S., Rabbi, M., Pelillo, M.: Content-based image retrieval with relevance feedback using random walks. *Pattern Recogn.* 44 (September 2011)

Author Index

- Ahlström, David I-81
Aizawa, Kiyoharu II-505
Anjyo, Ken II-110
Au, Oscar C. II-187
- Bai, Hongliang II-554
Bai, Wei I-403
Bailer, Werner I-81, I-456, II-385, II-538
Bao, Bing-Kun I-58, II-121
Barreda-Ángeles, Miguel I-456
Bloess, Mark I-130
Böszörmenyi, Laszlo II-88, II-544
Brenner, Markus I-185
- Cai, Haibin II-217
Cao, Liangliang II-133
Cao, Liujuan II-1, II-47
Cao, Yang I-272, II-12, II-154
Caprani, Niamh II-490
Chen, Fangdong II-511
Chen, Hong-Ming II-99
Chen, Yuanzhe I-163
Cheng, Jian II-417
Cheng, Wen-Huang II-99
Cheng, Zhiyong II-36
Chiu, Yung-Hsiang II-165
Cho, Joon-Myun II-479
Chu, Wei-Ta I-347
Chua, Tat-Seng I-70, I-141, II-514
Chung, Kuo-Liang II-165
Crane, M. I-490
Cui, Peng I-239
- Dai, Feng I-295, II-525, II-532
Deguchi, Daisuke II-364
Del Fabro, Manfred I-81, II-88, II-544
Deng, James J. I-524
Ding, Gangyi II-499
Ding, Jundi I-436
Dong, Yuan II-554
Duan, Lingyu II-143
Duan, Yizhou I-357
Duong, Duc Anh II-547
Dupont, Stéphane II-550
Dutoit, Thierry II-550
- El Saddik, Abdulmotaleb I-24, I-130
- Fan, Jianping II-194
Fang, Quan I-92
Fang, Shuai I-272
Fassold, Hannes II-385
Feng, Bailan I-250, II-374
Feng, Songhe II-428
Feng, Wengang I-152
Feng, Xiaoyi II-194
Frisson, Christian II-550
Fu, Haiyan II-293
- Gao, Ke I-261
Gao, Yue I-141, II-47, II-499
Gao, Zan I-206, II-517
Goh, Hai-Kiat II-502
Guan, Yue II-263, II-305
Guo, Dongyan I-436
Guo, Jinlin I-479, II-69, II-493, II-496, II-541
Guo, Junbo I-425, II-532
Guo, Yanqing II-293
Guo, Zongming I-357, I-403
Gurrin, Cathal I-479, I-490, II-69, II-490, II-493, II-496, II-541
- Han, Qi II-272
Hauptmann, Alexander II-499
He, Xuan II-508
Heng, Yue II-176
Heu, Jee-Uk II-479
Hofmann, Albert II-385
Hopfgartner, Frank I-479, II-69, II-493, II-541
Hou, Dejun II-514
Hsieh, Yung-Huan II-99
Hu, Min-Chun II-99
Hu, Wei II-187
Huang, Chun-Chang I-347
Huang, Qingming II-407
Huang, Yong-Huai II-165
- Ide, Ichiro II-58, II-364
Ikenaga, Takeshi I-336
Izquierdo, Ebroul I-185

- Jaimes, Alejandro I-535
 Jeong, Jin-Woo II-479
 Ji, Rongrong II-47
 Ji, Zhong I-217
 Jiang, Guang I-174
 Jiang, Linhua II-217
 Jiang, Shuqiang II-407
 Jiang, Tao II-528
 Jiang, Xiaolei I-368
 Jiao, Licheng I-283
 Joo, Young-Do II-479
 Jose, Joemon II-251
 Jung, Cheolkon I-283
- Kaliciak, Leszek I-445
 Kankanhalli, Mohan I-106
 Kar, Koushik II-334
 Katti, Harish I-106
 Kawamura, Soichiro II-505
 Kawate, Yuta II-110
 Kim, Heung-Nam I-24, I-130
 Kobayashi, Takashi II-364
 Kohara, Yuya I-47
 Kompatsiaris, Ioannis I-1
 Kong, Dehui II-468
 Kong, Xiangwei II-293
 Kruliš, Martin II-446
- Lam, Vu II-547
 Lang, Congyan II-428
 Lao, Songyang I-479
 Le, Duy-Dinh II-547
 Lee, Dong-Ho II-479
 Leong, Hon Wai I-318
 Leung, C.H.C. I-524
 Li, Bin II-511
 Li, Changsheng II-457
 Li, Gaojian I-163
 Li, Haojie II-206, II-263, II-305
 Li, Heping I-513
 Li, Houqiang II-511
 Li, Jiahong II-528
 Li, Jiaming II-521
 Li, Jing I-118
 Li, Jintao I-261, II-228, II-239, II-272,
 II-525
 Li, Lin I-318
 Li, Mading I-403
 Li, N. I-490
 Li, Qi II-79
- Li, Weifeng II-436
 Li, Yujun II-187
 Liang, Feidie II-239, II-272
 Liang, Feng II-535
 Liang, Ling-Ling II-316
 Liao, Qingmin I-414, II-436
 Liao, Weimin II-24
 Lin, Chih-Ming II-165
 Lin, Shouxun I-425
 Lin, Weiyao I-163
 Lin, Yan-Ching II-99
 Lipczak, Marek I-535
 Liu, Anan I-206
 Liu, Chaoteng I-118
 Liu, Cheng I-196
 Liu, Chih-Chin I-391
 Liu, Hong II-345
 Liu, Jiaying I-403
 Liu, Jing II-327
 Liu, Ju I-546
 Liu, Qinshan II-417
 Liu, Qiong II-1, II-47
 Liu, Shaowei I-239
 Liu, Xi I-174
 Liu, Xiangkai II-143
 Liu, Xiaocui I-546
 Liu, Yi I-425
 Liu, Zhenyu I-336
 Lokoč, Jakub II-446
 Lu, Hanqing II-327, II-417, II-457
 Lu, Jiayin II-293
 Lu, Ke I-92
 Lu, Shen II-24
 Luan, Huanbo I-239, II-514
 Luo, Suhuai II-521
- Ma, He I-380
 Mao, Wei II-502
 Mao, Xiaoyang I-467
 Martinez-Peñaranda, Carmen I-456
 McGuinness, Kevin II-541
 McParlane, Philip II-251
 Miao, Haiyan II-36
 Min, Weiqing II-121
 Mizui, Kenta I-502
 Moinet, Alexis II-550
 Morimoto, Masashi II-283
 Münzer, Bernd II-88, II-544
 Murase, Hiroshi II-364

- Neo, Shi-Yong II-206
 Ngo, Thanh Duc II-547
 Nguyen, Vu Hoang II-547
 Nie, Weizhi I-206

 O'Connor, Noel E. II-490, II-541
 Ohtani, Tomoko II-505
 Okabe, Makoto I-502, II-110
 Onai, Rikio I-502, II-110
 Ouyang, Zhe II-532

 Pan, Jeff I-445
 Pang, Yanwei I-217
 Papadopoulos, Symeon I-1
 Peng, Jinye II-194
 Peng, Qiang I-307, II-143, II-316
 Pereda-Baños, Alexandre I-456
 Picard-Limpens, Cécile II-550

 Qasim, Iqbal II-479
 Qi, Hongtao I-283
 Qian, Xueming I-118
 Qian, Yueliang II-345
 Qin, Zishan II-217
 Qu, Yanyun II-24

 Ramasubramonian, Adarsh K. II-334
 Ravet, Thierry II-550
 Rawashdeh, Majdi I-24, I-130
 Ren, Jing II-36
 Rong, Tie I-318
 Ruskin, H.J. I-490

 Sagonas, Christos I-1
 Sang, Jitao I-92
 Satoh, Shin'ichi II-547
 Sawada, Tomoya I-467
 Schallauer, Peter II-385
 Scherp, Ansgar I-36
 Schober, Christian II-538
 Schoeffmann, Klaus I-81
 Scott, David II-69, II-493, II-541
 Shao, Ling II-79
 Shen, Dingcheng II-517
 Shen, Heng Tao I-70
 Shen, Jialie II-36
 Shepherd, John I-228
 Shi, Zhongzhi I-13, I-174
 Siebert, Xavier II-550
 Skopal, Tomáš II-446

 Smeaton, Alan F. II-541
 Song, Dawei I-445
 Song, Xinhang II-407
 Sousa-Vieira, M.E. I-327
 Staab, Steffen I-36
 Su, Yunnyun II-217
 Su, Yuting I-206, I-217
 Sudo, Kyoko II-283
 Sun, Chao I-58
 Sun, Fuming II-263
 Sun, Jiande I-546
 Sun, Jun I-357, II-356
 Sun, Lei I-336
 Sun, Tao II-327
 Sun, Wenxiu II-187
 Sun, Xiaoshuai I-368
 Sun, Yongqing II-283

 Takahashi, Tomokazu II-364
 Tang, Jinhui I-436, II-407
 Tang, Sheng II-239, II-272
 Tang, Yuan Yan I-118
 Taniguchi, Yukinobu II-283
 Tao, Kun II-554
 Thallinger, Georg II-538
 Tian, Dongping I-13
 Tian, Qi I-239, II-228, II-395
 Tong, Wei II-499
 Toyoura, Masahiro I-467
 Tran, Vinh Quang II-547
 Trevisiol, Michele I-535
 Tu, Junnan II-217

 Vakali, Athena I-1

 Walber, Tina I-36
 Wang, Biao I-414, II-436
 Wang, Fangyuan I-513
 Wang, Hai I-513
 Wang, Hanli II-176
 Wang, Hongtao II-511
 Wang, Hongyi II-493
 Wang, Jing I-546
 Wang, Jinqiao II-457
 Wang, Lezi II-554
 Wang, Mei II-502
 Wang, Pengcheng II-468
 Wang, Shuhui II-407
 Wang, Wen II-428
 Wang, Xiangdong II-345

- Wang, Xiaotao II-535
 Wang, Xu II-1
 Wang, Yang I-228, II-133
 Wang, Yu II-239, II-272
 Wang, Zengfu II-12, II-154
 Wechtitsch, Stefanie II-385
 Wei, Hui I-163
 Weiss, Wolfgang I-81, I-456, II-538
 Whiting, Stewart II-251
 Wiratunga, Nirmalie I-445
 Woods, John W. II-334
 Wu, Jianxin I-163
 Wu, Lin I-228
 Wu, Meng II-356
 Wu, Rihui II-528
 Wu, Shaojie II-24
 Wu, Xiao I-307, II-143, II-316

 Xia, Xu II-143
 Xia, Zhaoqiang II-194
 Xiang, Kang II-528
 Xiao, Bo II-176
 Xie, Yaoqin II-79
 Xu, Bo I-250, II-374
 Xu, Changsheng I-58, I-92, II-121
 Xu, Guangping II-517
 Xu, Lingfeng II-187
 Xu, Pengfei I-368
 Xu, Su II-374
 Xu, Tao II-345
 Xue, Feng II-535
 Xue, Yanbing II-517

 Yadati, Karthik I-106
 Yan, Chenggang I-295
 Yanai, Keiji I-47
 Yang, Linjun I-118
 Yang, Pingping II-535
 Yang, Shiqiang I-239
 Yang, Wei-Ning II-165
 Yang, Wenming I-414
 Yang, Yang I-70, II-493, II-541
 Yang, You II-1
 Yao, Hongxun I-368
 Yi, Lei II-206, II-305
 Yin, Baocai II-468
 Yu, Jen-Yu I-347
 Yu, Jun II-12

 Yu, Xinguo I-318
 Yu, Yanru I-217
 Yuan, Hui I-546
 Yuan, Ting II-417
 Yuan, Zhe I-272
 Yue, JinPeng I-174

 Zha, Zheng-Jun I-70
 Zhang, Bo I-250
 Zhang, Chunjie II-327
 Zhang, Dong II-511
 Zhang, Dongming I-261, I-425, II-228,
 II-395, II-525
 Zhang, Hang II-334
 Zhang, Hao II-305, II-508
 Zhang, Hua II-517
 Zhang, Jing I-272, II-154
 Zhang, Jun I-295
 Zhang, Lei I-307, II-143, II-395
 Zhang, Longfei II-499
 Zhang, Sheng I-163
 Zhang, Shijie I-272
 Zhang, Shuwu I-513
 Zhang, Wei I-261
 Zhang, Xi II-417
 Zhang, YaLin II-239
 Zhang, Ying I-380
 Zhang, Yong II-468
 Zhang, Yongdong I-295, II-395, II-532
 Zhang, Zhenxing I-479, II-493, II-496,
 II-541
 Zhao, Bo II-316
 Zhao, Chunxia I-436
 Zhao, Sicheng I-368
 Zhao, Tiesong II-176
 Zhao, Xiaofei I-13
 Zhao, Zhicheng I-196
 Zheng, Zhigang II-154
 Zhou, Fei I-414
 Zhou, Jun II-356
 Zhou, Lijuan Marissa II-490, II-493
 Zhu, Guibo II-457
 Zhu, Qingsong II-79
 Zhu, Wenwu I-239
 Zhu, Xiao Yu II-508
 Zhuang, Dongye II-228
 Zimmermann, Roger I-380